Blending Fast and Slow Thinking: Unlocking the Power of Multimodal Large Language Models

Xin Cai

Independent Researcher
xincai00@gmail.com

1 Introduction

Large Language Models (LLMs) [61] have achieved remarkable milestones, especially the demonstrated emergent capabilities [48], e.g., In-Context Learning (ICL), instruction following, and Chain of Thought (CoT) reasoning [49]. It is now an inevitable trend to integrate various modalities with LLMs to mimic the way humans interact with the open world, thus ushering in the new era of Multimodal Large Language Models (MLLMs[1]) [55].

Recent work [24] has demonstrated that the development of MLLMs signifies a transition from specialist models to general-purpose visual assistants, regarded as a crucial step towards realizing the grand vision of a general-purpose multimodal AI agent. It is further unveiled in [24] that a converging point for many development trajectories of vision-language/multimodal foundation models is *"the creation of general-purpose models and systems capable of following human intents and effortlessly executing a diverse array of vision and vision-language tasks in the wild."*

The key features that markedly differentiate MLLMs from deep learning models developed beforehand are condensed into the name itself: General Purpose Assistants, which include **versatility**, **interactivity**, and **controllability**. Specifically, the quest for a unified architecture capable of accomplishing a diverse range of discriminative and generative tasks has witnessed great success in Natural Language Processing (NLP), as exemplified by GPT-3, inspiring similar research endeavours in the field of computer vision. However, as pointed out in [24], the development of unified vision systems significantly lags behind due to fragmented vision tasks and the difficulty of scaling up visual data. Furthermore, to enable seamless communication between machines and humans, natural language serves as a core component in building a general interactive interface, which can be further complemented by multimodal prompts such as clicks, scribbles, or voice commands provided by users. Additionally, researchers have made strides in steering the output behaviour of MLLMs by instruction tuning or aligning with human preferences, marking a hallmark in elevating machine intelligence to the next level as humans begin to positively intervene in the machine

[1] Please note that the term MLLMs hereafter is used interchangeably with Large Vision-Language Models (LVLMs) or Large Multimodal Models (LMMs) given that the transitive property of modality embeddings showcased in ImageBind [11].

learning process. By leveraging human feedback and rich experience, machines can evolve more effectively and efficiently, leading to better outcomes for both machines and humans.

2 Technical Overview of MLLMs

The recent progress of MLLMs [2] can be roughly divided into the following three aspects: 1) architectures, 2) training strategies and data, 3) evaluation benchmarks. [3]

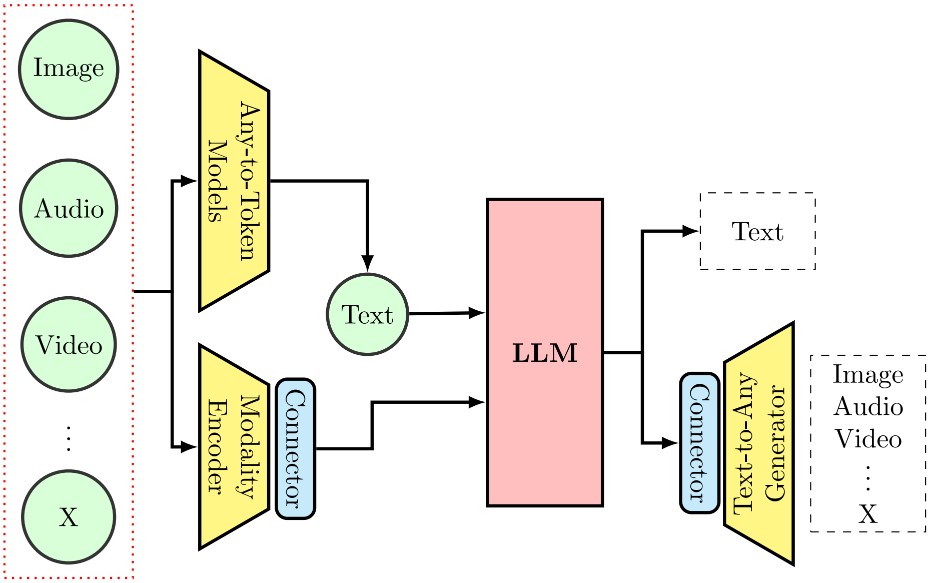

Fig. 1: An overview of MLLM architecture.

2.1 Architectures

A typical neural architecture of MLLMs consists of pre-trained modality encoders, pre-trained LLMs, learnable modality adapters/connectors, and optional modality decoders.

[2] Technically, the progress here excludes that related to Multimodal Agents, i.e., chaining tools with LLMs.

[3] For a more detailed introduction, please refer to my blog post.

Modality Encoders When compressing raw images, it is common practice to utilize pre-trained CLIP [37] image encoders or other similar variants [42], leveraging a pre-aligned albeit coarse vision-semantic space. The underlying principle is to discretize/standardize modality embeddings with reference to language vocabulary, which can be extended to encoding other kinds of multimodal signals.

LLMs Regarding pre-trained LLMs, open-source models such as the LLaMA series [44, 45] and Vicuna family [7] have gained widespread popularity in academic research. These models are often combined with Parameter-Efficient Fine-Tuning (PEFT) techniques [14], such as Low-Rank Adaptation (LoRA) [18], in the Supervised Fine-Tuning (SFT) phase to enhance instruction following and alignment with human preference. Furthermore, LLMs constructed with a Mixture of Experts (MoE) [21, 40] have attracted increasing attention due to the reduced computational cost of the sparse architecture.

Modality Connectors Due to the noticeable gap between embeddings of natural languages and other modalities, particularly considering the difference in bandwidth, it is essential to establish reliable alignment with reference to textual embeddings to enable LLMs to understand sensory inputs. This alignment involves projecting embeddings from modality encoders into a space comprehensible to LLMs, ensuring unambiguous interpretation. This can be achieved using a trainable modality adapter or connector, bridging the frozen modality encoder and the LLM, which is computationally efficient and more accessible.

Modality Decoders In addition to the aforementioned three compulsory modules, MLLMs can be further extended with modality-specific decoders, particularly in the Any-to-Any workflow, to equip MLLMs with multimodal responsive capabilities.

2.2 Training Strategies and Data

The training process of MLLMs can be decomposed into three stages: i) pre-alignment, ii) instruction tuning, iii) alignment with human preferences. The fulfilment of training objective in each phase heavily relies on specific data, necessitating cost-effective data scaling-up methods and high quality data.

Pre-Alignment The pre-alignment stage often requires an enormous amount of text-paired data, such as images, videos, or audio files with associated textual descriptions, usually gathered from the Internet.

Instruction Tuning Instruction tuning (IT) [60] aims to steer LLMs to respond more faithfully to human instructions by fine-tuning on instruction following datasets, which consist of (Instruction, Input, Output) triplets. This technique has proven to be effective and computationally efficient in enhancing the controllability of LLMs' output behaviour, thereby eliciting knowledge from LLMs that is well-aligned with human intents. Furthermore, IT has been identified as a critical factor in unlocking the few-shot (ICL) or zero-shot generalization capability of LLMs on solving novel tasks.

Preference Alignment Alignment with human preferences aims to further refine the output behaviour of MLLMs that have undergone SFT or instruction tuning by leveraging human/AI feedback on model responses. The two mainstream solutions currently in use are Reinforcement Learning with Human Feedback (RLHF) [35, 63] and Direct Preference Optimization (DPO) [38].

2.3 Evaluation Benchmarks

The evaluation of MLLMs is generally categorized into closed-ended and open-ended questions. For open-ended questions, the assessment criteria may involve human or GPT scoring. In contrast to evaluation methods developed before the era of LLMs, designing evaluation toolkits for MLLMs demands attention and efforts comparable to or even surpass those required for model development. As pointed out in [12], human evaluators without domain expertise or significant time investment can be easily deceived by LLMs due to their fluent, confident, and well-structured responses. Therefore, acquiring quantitative measurements that can objectively and reliably reflect the multifaceted aspects of MLLMs is crucial, as these measurements facilitate pinpointing critical factors contributing to improvements along specified skill dimensions, identifying failure cases, and providing deeper insights into the inner workings of MLLMs.

3 The Outline of Research Objectives and Methodologies

My overarching research objective is to gain deeper insights into the inner mechanisms that govern the behaviour of MLLMs, which can be decomposed into mitigating multimodal hallucinations, investigating the reasoning capabilities of LLMs when extended to the multimodal setting, and building universal video understanding and reasoning systems aided by LLMs. Additionally, current (M-)LLMs face inherent limitations due to its purely data-driven nature, such as the inability to access up-to-date information and leverage specialized modules, the single-step question-to-answer inference scheme that hinders interpretability, and the prohibitive computational demands on compute and data. To overcome these limitations, it is promising to investigate techniques complementary to connectionist methods (e.g., deep learning), such as neural-symbolic learning, model predictive control (MPC) and reinforcement learning (RL).

3.1 Multimodal Hallucination

Multimodal Hallucination refers to MLLMs generating responses inconsistent with the concrete content present in multimodal signals. Currently, it is primarily constrained to object-centred hallucination, which includes the existence of objects, erroneous attribute binding, and spurious relationships between objects. This phenomenon reflects the inherent difficulty of building a well-aligned and harmonious multimodal embedding space, especially when integrated into a pre-established language embedding space of LLMs.

The work POPE [27] presents the first systematic investigation of object hallucination in Large Vision Language Models (LVLMs). It reveals that object hallucination is intimately related to the frequency of objects appearing in visual instruction tuning data, or the frequency of co-occurring objects, referred to as statistical bias in the work VCD [23]. Additionally, VCD [23] points out that another contributing factor is the over-reliance on a strong language prior, leading to the oversight of visual content, referred to as parametric knowledge in the work HallE-Control/-Switch [58]. Despite these pioneering efforts, the deep causes of multimodal/object hallucination and corresponding reliable evaluation methods remain elusive. On the contrary, it has been noted that current state-of-the-art text-to-image generation models, e.g., diffusion models, struggle to faithfully produce content adhering to the semantic meanings in the text prompts, referred to as catastrophic forgetting in the work [3]. In other words, balancing the generation diversity and faithfulness of the generated content will be a long-standing challenge in MLLMs research.

The mitigating strategies for multimodal hallucination can be divided into three categories: i) pre-correction, ii) in-process correction, and iii) post-correction.

Pre-correction leverages negative/hallucinated responses during the instruction or alignment tuning stage to correct MLLMs' output behaviour [29].

Another line of research, called in-process correction, addresses this issue by investigating plausible causes of hallucination throughout the entire pipeline, including but not limited to examining constituent modules, Supervised Fine-Tuning (SFT) data, and image resolution. Inspired by contrastive decoding [26], existing works propose to recalibrate the output probability distribution of LLMs used for beam-search decoding with reference to unconditional or ungrounded generation [23, 58].

The last type of solutions, known as post-correction, aims to alleviate object hallucination by using a diagnosis-then-rectify paradigm to scrutinize the output responses of MLLMs. For example, Woodpecker [56] breaks down generated responses potentially containing hallucinated answers into five explicit stages. Semantic concepts with visual correspondence are double-checked by expert vision perception models, such as open-vocabulary detectors [30]. These models serve as a complementary visual knowledge base, which is subsequently combined with the original prompts fed into MLLMs again to remedy hallucination.

3.2 Multimodal Reasoning

Reasoning is the cognitive process by which the human brain integrates heterogeneous sources of information, evaluates their validity and relevance, and draws conclusions to make informed decisions or solve problems effectively.

Research on the reasoning capabilities of LLMs is still in its infancy and lacks a rigorous, unanimous definition of the concept of reasoning. Generally, it is referred to as the exploitation of world knowledge embedded in LLMs or the ability to tackle complex tasks requiring compositional generalization [22]. It has been demonstrated that CoT prompting [49] or fine-tuning on CoT data [20] can significantly improve the performance of LLMs on various reasoning tasks [8, 31, 33] due to the enriched context where outcomes produced from intermediate reasoning steps (i.e., rationales) facilitate the derivation of correct answers. The divide-and-conquer strategy is also commonly employed to decompose a complex problem into more manageable subproblems [62]. Furthermore, LLMs capable of program synthesis [5] have been shown to perform reasonably well on reasoning tasks [6, 10], possibly benefiting from a better understanding of logical operators and control flows in high-level programming languages such as Python.

While it remains controversial whether LLMs are actually capable of reasoning [46], there has been a surge of interest in harnessing their emergent capabilities to enhance multimodal understanding and reasoning[4], particularly in the vision domain. The extension of the reasoning ability of LLMs beyond the text modality, however, faces additional challenges. Inspired by Neural Module Networks [1, 2], where visual reasoning is argued to be inherently compositional and thus necessitates a modularity design principle, a line of research revolves around capitalizing on external tools or expert models with an LLM functioning as a coordinator or planner [13, 32, 41, 43, 50, 53], which enables effective communication among diverse specialized modules in a natural language format. This modularity design principle is also firmly grounded in cognitive neuroscience theory. Specifically, the dual-system theory framework [9] posits the existence of two distinct systems that guide human cognitive activities. System I (the fast mode of thinking) can be modelled with end-to-end optimized deep neural networks for multimodal perception or pattern recognition, while System II (the slow mode of thinking) is characterized by deliberate, analytical thinking processes and demands disparate modelling paradigms with particular requirements for explainability and interpretability.

When constrained to multimodal reasoning, inspired by the modularity of cognitive functions, it is essential to deliberately separate the visual processing and reasoning responsibilities of MLLMs rather than pursuing a monolithic architecture competent in all skill dimensions, which requires exploring effective modelling paradigms for System II and the integration of Systems I and II. Specifically, Visual Programmatic Models (VPMs) [13, 43] instantiate the principle of deliberate separation of visual processing and reasoning responsibilities of MLLMs, leveraging a code LLM (e.g., Codex [5]) to synthesize customized programs for each query, where various tools (e.g., LLMs, vision expert models, Python built-in functions, web search engines) are composed following control flows (e.g., if-else statements, for loops) to accomplish complex visual reasoning tasks. VPMs can be regarded as a specific type of tool-augmented LLMs, with differences in how scripts of tool invocations are composed—whether through naive word matching [53], high-level programming languages [13, 43], or plain natural language [32, 41, 50].

In addition to tool-augmented LLMs, other works have focused on exploring effective visual CoT strategies [4, 16, 28, 34, 36, 39, 51, 57]. The foremost challenge of visual CoT is the identification of visual clues essential for addressing in-

[4] Multimodal reasoning will hereafter be used interchangeably with visual reasoning, as the current focus is primarily on visual signals, such as images or videos.

put queries [4,28,36,39,51], i.e., grounding visual concepts relevant to language queries, which also facilitates mitigating multimodal hallucinations of MLLMs. Recent work highlights the significance of incorporating efficient visual search mechanisms [28,39,51] into MLLMs, with precisely located visual elements serving as reasoning anchors in visual CoT, advocating for the object-centric decomposition of complex visual scenes [28]. Furthermore, it has been demonstrated that structured representations of visual scenes [16, 34], such as scene graphs where relationships among interconnected objects are specified, provide richer contexts in CoT for compositional visual reasoning. Regarding video understanding, the visual reasoning capability of MLLMs is contingent on the fine-grained temporal grounding of instances and events [36]. Apart from adopting visual CoT for prompting [4, 34, 57], it can also be employed for instruction tuning of MLLMs [16, 28, 36, 39, 51]. However, the cost of curating large-scale visual CoT data with accurately annotated intermediate outcomes is prohibitive for manual efforts, necessitating scalable (semi-)automatic data engines [19, 36] to expedite the procedure. Specifically, Visual Program Distillation (VPD) [19] proposed to convert executable programs invoking specialized tools to solve each reasoning step explicitly to natural language CoT instruction data, which are then distilled [17] into a VLM, demonstrating a promising solution to integrate System I and II.

Drawing inspiration from the computational models [51, 54] for mimicking humans' visual search process, developing effective methods to enhance the reasoning capabilities of MLLMs can be facilitated by a deeper understanding of the reasoning processes in the human brain. Key brain regions involved in reasoning include the prefrontal cortex, which is critical for higher-order cognitive functions; the parietal cortex, responsible for integrating sensory information and abstract reasoning; and the hippocampus, essential for memory formation and retrieval. By drawing parallels between these brain regions and the architectural components of MLLMs, we can develop more sophisticated and explainable models that better emulate human reasoning.

3.3 MLLMs for Video Understanding

The next milestone in the evolution of existing image-related multimodal dialogue systems is marked by conquering universal video understanding powered by LLMs. The added temporal dimension in videos significantly amplifies the complexities that need to be addressed by MLLMs. These include intricate spatio-temporal understanding and reasoning, causal effects, the synergy between visual, textual, and auditory signals, inherited and exacerbated hallucination, limited context length, prohibitive training and inference costs, scalable methods for constructing high-quality instruction tuning data, and reliable evaluation benchmarks.

Pioneer attempts to craft video-centric MLLMs can be divided into two categories: i) verbalizing video signals aided by various Foundation Models [25,47], and ii) building an end-to-end video understanding system [25,59] by following a similar pattern to image-based MLLMs: vision encoders → modality connectors

→ LLMs. For example, the work ChatVideo [47] constructs a tracklet-centric database storing various attributes grouped based on unique identifiers of tracklets enabled by diverse Video Foundation Models (ViFMs). This database is subsequently queried by ChatGPT to respond to user instructions by grounding on relevant video content. A similar idea has been explored in the VideoChat-Text model [25].

The first line of research features a divide-and-conquer strategy and may benefit from transparency and interpretability due to the decomposed tasks. However, it suffers from information loss during the conversion from raw but rich video signals to highly symbolic text descriptions and from the prohibitive cost of invoking ViFMs.

The second line of research aims to build an end-to-end video understanding and reasoning pipeline with LLMs as the nexus, exemplified by VideoChat-Embed [25] and Video-LLaMA [59]. However, the absence of One-For-All ViFMs [47], unifying a broad array of video understanding tasks such as video object tracking, video captioning or QA, and temporal event localization or grounding, casts further doubt on the feasibility of a single but omnipotent video encoder. Even without viability concerns, the realization of spatio-temporal understanding and reasoning places high demands on the scale and quality of instruction tuning data, the difficulty level of which would be escalated with the increasingly refined granularity in either spatial or temporal dimensions. Furthermore, multimodal hallucination would be exacerbated as incompetent video encoders and flawed training recipes can amplify the tendency of MLLMs to exploit parametric knowledge obtained from massive pre-trained language corpora. This is evidenced by a recent comprehensive study of MLLMs' capabilities and limitations in VideoQA [52], where it has been shown that MLLMs do not possess a clear edge over non-LLM techniques in terms of temporal understanding, visual grounding, multimodal reasoning, robustness, and generalization. Last but not least, long-term video understanding ($\geq$ 1 minute) [15] remains a longstanding challenge that persistently plagues the research community, necessitating a meticulous balance between effectiveness and efficiency by expanding considerations to cover context length, GPU memory usage, response latency, and even hardware-related factors.

References

1. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 39–48 (2016)
2. Assouel, R., Rodriguez, P., Taslakian, P., Vazquez, D., Bengio, Y.: Oc-nmn: Object-centric compositional neural module network for generative visual analogical reasoning. arXiv preprint arXiv:2310.18807 (2023)
3. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)

4. Chen, J., Liu, Y., Li, D., An, X., Feng, Z., Zhao, Y., Xie, Y.: Plug-and-play grounding of reasoning in multimodal large language models. arXiv preprint arXiv:2403.19322 (2024)
5. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)
6. Chen, W., Ma, X., Wang, X., Cohen, W.W.: Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. arXiv preprint arXiv:2211.12588 (2022)
7. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) **2**(3), 6 (2023)
8. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457 (2018)
9. Daniel, K.: Thinking, fast and slow (2017)
10. Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., Neubig, G.: Pal: Program-aided language models. In: International Conference on Machine Learning. pp. 10764–10799. PMLR (2023)
11. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023)
12. Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., Song, D.: The false promise of imitating proprietary llms. arXiv preprint arXiv:2305.15717 (2023)
13. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14953–14962 (2023)
14. Han, Z., Gao, C., Liu, J., Zhang, S.Q., et al.: Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608 (2024)
15. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. arXiv preprint arXiv:2404.05726 (2024)
16. Herzig, R., Mendelson, A., Karlinsky, L., Arbelle, A., Feris, R., Darrell, T., Globerson, A.: Incorporating structured representations into pretrained vision & language models using scene graphs. arXiv preprint arXiv:2305.06343 (2023)
17. Hsieh, C.Y., Li, C.L., Yeh, C.K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. arXiv preprint arXiv:2305.02301 (2023)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
19. Hu, Y., Stretcu, O., Lu, C.T., Viswanathan, K., Hata, K., Luo, E., Krishna, R., Fuxman, A.: Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9590–9601 (2024)
20. Huang, J., Gu, S.S., Hou, L., Wu, Y., Wang, X., Yu, H., Han, J.: Large language models can self-improve. arXiv preprint arXiv:2210.11610 (2022)

21. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
22. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40**, e253 (2017)
23. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. arXiv preprint arXiv:2311.16922 (2023)
24. Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J.: Multimodal foundation models: From specialists to general-purpose assistants. arXiv preprint arXiv:2309.10020 **1**(2), 2 (2023)
25. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
26. Li, X.L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., Lewis, M.: Contrastive decoding: Open-ended text generation as optimization. arXiv preprint arXiv:2210.15097 (2022)
27. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
28. Li, Z., Luo, R., Zhang, J., Qiu, M., Wei, Z.: Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. arXiv preprint arXiv:2405.16919 (2024)
29. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: The Twelfth International Conference on Learning Representations (2023)
30. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
31. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
32. Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.W., Wu, Y.N., Zhu, S.C., Gao, J.: Chameleon: Plug-and-play compositional reasoning with large language models. Advances in Neural Information Processing Systems **36** (2024)
33. Lu, P., Qiu, L., Chang, K.W., Wu, Y.N., Zhu, S.C., Rajpurohit, T., Clark, P., Kalyan, A.: Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. arXiv preprint arXiv:2209.14610 (2022)
34. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14431 (2024)
35. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
36. Qian, L., Li, J., Wu, Y., Ye, Y., Fei, H., Chua, T.S., Zhuang, Y., Tang, S.: Momentor: Advancing video large language model with fine-grained temporal reasoning. arXiv preprint arXiv:2402.11435 (2024)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

38. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36** (2024)
39. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. arXiv preprint arXiv:2403.16999 (2024)
40. Shen, S., Hou, L., Zhou, Y., Du, N., Longpre, S., Wei, J., Chung, H.W., Zoph, B., Fedus, W., Chen, X., et al.: Mixture-of-experts meets instruction tuning: A winning combination for large language models. arXiv preprint arXiv:2305.14705 (2023)
41. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. Advances in Neural Information Processing Systems **36** (2024)
42. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
43. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11888–11898 (2023)
44. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
45. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
46. Valmeekam, K., Olmo, A., Sreedharan, S., Kambhampati, S.: Large language models still can't plan (a benchmark for llms on planning and reasoning about change). arXiv preprint arXiv:2206.10498 (2022)
47. Wang, J., Chen, D., Luo, C., Dai, X., Yuan, L., Wu, Z., Jiang, Y.G.: Chatvideo: A tracklet-centric multimodal and versatile video understanding system. arXiv preprint arXiv:2304.14407 (2023)
48. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
49. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
50. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
51. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13084–13094 (2024)
52. Xiao, J., Huang, N., Qin, H., Li, D., Li, Y., Zhu, F., Tao, Z., Yu, J., Lin, L., Chua, T.S., et al.: Videoqa in the era of llms: An empirical study. arXiv preprint arXiv:2408.04223 (2024)
53. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023)

54. Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., Hoai, M.: Predicting goal-directed human attention using inverse reinforcement learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 193–202 (2020)
55. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. arXiv preprint arXiv:2306.13549 (2023)
56. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. arXiv preprint arXiv:2310.16045 (2023)
57. You, H., Sun, R., Wang, Z., Chen, L., Wang, G., Ayyubi, H.A., Chang, K.W., Chang, S.F.: Idealgpt: Iteratively decomposing vision and language reasoning via large language models. arXiv preprint arXiv:2305.14985 (2023)
58. Zhai, B., Yang, S., Zhao, X., Xu, C., Shen, S., Zhao, D., Keutzer, K., Li, M., Yan, T., Fan, X.: Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. arXiv preprint arXiv:2310.01779 (2023)
59. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
60. Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al.: Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792 (2023)
61. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)
62. Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al.: Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625 (2022)
63. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)