

COMPUTER SCIENCE



A Survey on Multimodal Large Language Models

Shukang Yin^{1,†}, Chaoyou Fu^{2,3,*,†}, Sirui Zhao^{1,*,†}, Ke Li⁴, Xing Sun⁴, Tong Xu¹ and Enhong Chen^{1,*}

ABSTRACT

Recently, Multimodal Large Language Model (MLLM) represented by GPT-4V has been a new rising research hotspot, which uses powerful Large Language Models (LLMs) as a brain to perform multimodal tasks. The surprising emergent capabilities of MLLM, such as writing stories based on images and Optical Character Recognition (OCR)-free math reasoning, are rare in traditional multimodal methods, suggesting a potential path to artificial general intelligence. To this end, both academia and industry have endeavored to develop MLLMs that can compete with or even outperform GPT-4V, pushing the limit of research at a surprising speed. In this paper, we aim to trace and summarize the recent progress of MLLMs. First of all, we present the basic formulation of MLLM and delineate its related concepts, including architecture, training strategy and data, as well as evaluation. Then, we introduce research topics about how MLLMs can be extended to support more granularity, modalities, languages, and scenarios. Finally, we continue with multimodal hallucination and extended techniques, including Multimodal ICL (M-ICL), Multimodal CoT (M-CoT), and LLM-Aided Visual Reasoning (LAVR). To conclude the paper, we discuss existing challenges and point out promising research directions.

Keywords: multimodal large language model, vision language model, large language model

INTRODUCTION

In recent years, there have been remarkable progress in Large Language Models (LLMs) [1,2]. As scaling up data size and model size, these LLMs raise extraordinary emergent abilities, typically including instruction following [3], In-Context Learning (ICL) [4], and Chain of Thought (CoT) [5]. Although LLMs have demonstrated surprising zero/few-shot reasoning performance on most Natural Language Processing (NLP) tasks [6] and even complex real-life applications [7–9], they are inherently “blind” to vision since they can only understand discrete text. At the same time, Large Vision Models (LVMs) can see clearly [10,11], but commonly lag in reasoning.

Out of this complementarity, LLM and LVM run towards each other, leading to the new field of Multimodal Large Language Model (MLLM). Typically, it refers to the LLM-based model with the ability to receive, reason, and output with multimodal information. Prior to MLLM, there have been a lot of works devoted to multimodality, which can be divided into discriminative [12,13] and generative [14,15]

paradigms. QLM [12], as a representative of the former, projects visual and textual information into a unified representation space, building a bridge for downstream multimodal tasks. In contrast, OFA [14] is a representative of the latter, which unifies multimodal tasks in a sequence-to-sequence manner. MLLM can be classified as the latter according to the sequence operation, but it manifests two distinct traits compared with its traditional counterparts: (1) MLLM is based on LLM with billion-scale parameters, which is not available in previous models. (2) MLLM uses new training paradigms to unleash its full potential, such as using multimodal instruction tuning [16] to encourage the model to follow new instructions. Equipped with the two traits, MLLM exhibits new capabilities, such as writing website code based on images [17], understanding the deep meaning of a meme [18], and OCR-free math reasoning [19].

Ever since the release of GPT-4 [20], there has been a research frenzy over MLLMs because of the amazing multimodal examples it shows. This development is fueled by ef-

¹University of Science and Technology of China, Hefei 230026, China;

²State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China;

³School of Intelligence Science and Technology, Nanjing University, Suzhou 215163, China;

⁴Tencent YouTu Lab, Shanghai 200233, China

*Corresponding authors.

Email:
bradyfu24@gmail.com;
sirui@mail.ustc.edu.cn;
cheneh@ustc.edu.cn.

[†]Equally contributed to this work.

Received: XX XX Year;
Revised: XX XX Year;
Accepted: XX XX Year

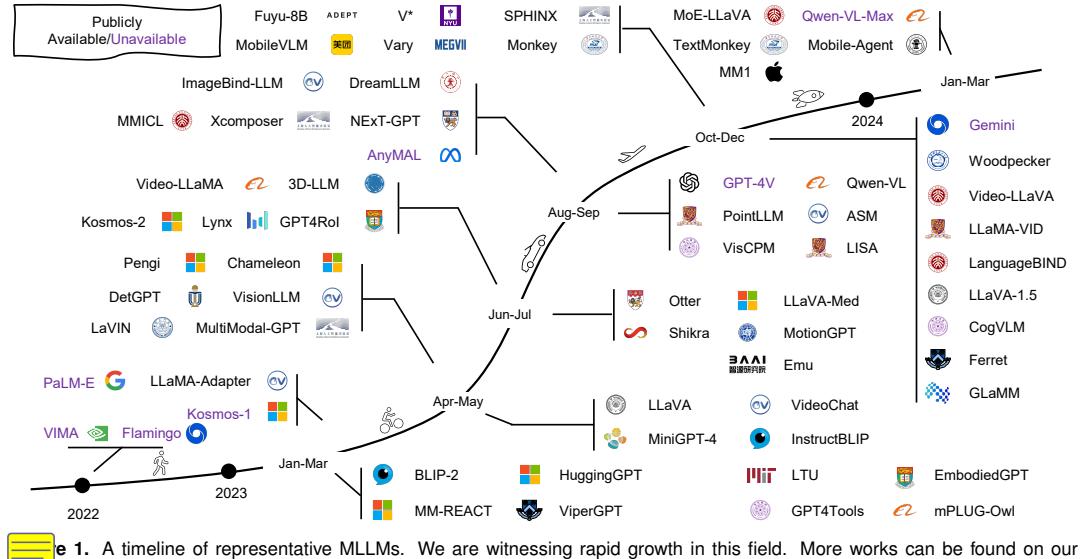


Fig. 1. A timeline of representative MLLMs. We are witnessing rapid growth in this field. More works can be found on our released GitHub page, which is updated daily.

forts from both academia and industry. Preliminary research on MLLMs focuses on text content generation grounded in text prompts and image [16]/video [21,22]/audio [23]. Subsequent works have expanded the capabilities or the usage scenarios, including: Better granularity support. Finer control on user prompts is developed to support specifying regions through boxes [24] or a certain object through a click [25]. Enhanced support on input and output modalities [26,27], such as image, video, audio, and point cloud. Improved language support. Efforts have been made to extend the success of MLLMs to other languages (*e.g.* Chinese) with relatively limited training corpus [28]. Extension to more realms and usage scenarios. Some studies transfer the strong capabilities of MLLMs to other domains, such as medical image understanding [29] and document parsing [30]. Moreover, multimodal agents are developed to assist in real-world interaction, *e.g.* embodied agents [31] and GUI agents [32]. MLLM timeline is illustrated in Fig. 1.

View of such rapid progress and the promising results of this field, we write this survey to provide researchers with a grasp of the basic idea, main method, and current progress of MLLMs. Note that we mainly focus on visual and language modalities, but also include works involving other modalities like video and audio. Specifically, we cover the most important aspects of MLLMs with corresponding summaries and open a GitHub page that would be updated in real time. To the best of our knowledge, this is the first survey on MLLMs.

following parts of the survey are struc-

tured as such: the survey starts with a comprehensive review of the essential aspects of MLLMs, including Mainstream architecture; (2) A full recipe of training strategy and data; (3) Common practices of performance evaluation. Then, we delve into a deeper discussion on some important topics about MLLMs, each focusing on a main problem: What aspects can be further improved or extended? How to relieve the multimodal hallucination issue? survey continues with the introduction of three key techniques, each specialized in a specific scenario: CL is an effective technique commonly used at the inference stage to boost few-shot performance. Other important technique is M-CoT, which is typically used in complex reasoning tasks. afterward, we delineate general ideas for developing LLM-based systems to solve composite reasoning tasks or to address common user queries. finish our survey with a summary and potential research directions.

ARCHITECTURE

Typical MLLM can be abstracted into three modules, *i.e.* a pre-trained modality encoder, a pre-trained LLM, and a modality interface to connect them. Using an analogy to humans, modality encoders such as image/audio encoders are human eyes/ears that receive and pre-process optical/acoustic signals, while LLMs are like human brains that understand and reason with the processed signals. Between, the modality interface serves to align different modalities. The MLLMs also include a generator to output other modalities apart from text. Diagram of the architecture is plotted in Fig. 2. In this section, we

introduce each module in sequence.

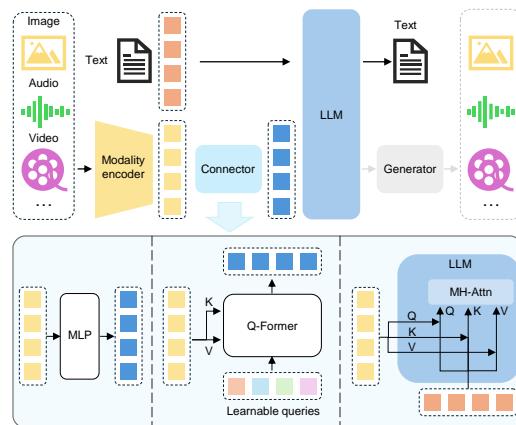


Figure 2. An illustration of typical MLLM architecture. It includes an encoder, a connector, and a LLM. A conditional generator can be attached to the LLM to generate more modalities besides text. The encoder takes in images, audios or videos and outputs features, which are processed by the connector so that the LLM can better understand. There are broadly three types of connectors: projection-based, query-based, and fusion-based connectors. The former two types adopt token-level fusion, processing features into tokens to be sent along with text tokens, while the last type enables a feature-level fusion inside the LLM.

Modality encoder

Encoders compress raw information, such as images or audio, into a more compact representation. Rather than training from scratch, a common approach is to use a pre-trained encoder that has been aligned to other modalities. For example, CLIP [12] incorporates a visual encoder semantically aligned with the text through large-scale pre-training on image-text pairs. Therefore, it is more practical to utilize such pre-aligned encoders to align with LLMs through alignment pre-training.

Commonly used image encoders are summarized in Table 1. Apart from vanilla CLIP image encoders [12], some works also explore using other variants. For example, MiniGPT-4 [17] adopts an EVA-CLIP [36] (ViT-G/14) encoder, which is trained with improved training techniques. Grey [25] introduces a convolution-based ConvNext-L encoder [33] to utilize higher resolution and multi-level features. These works also explore encoder-free architecture. For instance, the image patches of Fuyu-8b [37] are directly projected before sending to LLMs. In this design, the model naturally supports flexible input image resolution.

In choosing encoders, one often considers factors like resolution, parameter size, and pretraining corpus. Typically, many works have

empirically verified that using higher resolution can achieve remarkable performance gains [28, 38]. Approaches for scaling up input resolution can be categorized into direct scaling and patch-division methods. Direct scaling way inputs images of higher resolutions to the encoder, which often involves further tuning the encoder [28] or replacing a pre-trained encoder with higher resolution [39]. Similarly, CogAgent [32] uses a dual-encoder mechanism, where two encoders process high and low-resolution images, respectively. High-resolution features are injected into the low-resolution branch through cross-attention. Patch-division methods cut a high-resolution image into patches and reuse the low-resolution encoder. For example, Monkey [38] and SPHINX [40] divide a large image into smaller patches and send sub-images together with a downsampled high-resolution image to the image encoder, where the sub-images and the low-resolution image capture local and global features, respectively. In contrast, parameter size and training data composition are of less importance compared with input resolution, found by empirical studies [41].

Similar encoders are also available for other modalities. For example, Pengi [23] uses CLAP [42] model as the audio encoder. ImageBind-LLM [26] uses the ImageBind [43] encoder, which supports encoding image, text, audio, depth, thermal, and Inertial Measurement Unit (IMU) data. Equipped with the strong encoder, ImageBind-LLM can respond to the input of various modalities.

Pre-trained LLM

Instead of training an LLM from scratch, it is more efficient and practical to start with a pre-trained one. Through tremendous pre-training on web corpus, LLMs have been embedded with rich world knowledge, and demonstrate strong generalization and reasoning capabilities.

We summarize the commonly used and publicly available LLMs in Table 2. Notably, most LLMs fall in the causal decoder category, following GPT-3 [4]. Among them, Flan-T5 [44] series are relatively early LLMs used in works like BLIP-2 [50] and InstructBLIP [51]. The MA series [45] and Vicuna family [46] are representative open-sourced LLMs that have attracted much academic attention. Since the two LLMs are mainly pre-trained on English corpus, they are limited in multi-language support, such as Chinese. In contrast, Qwen [48] is a bilingual LLM with Chinese and English support.

It should be noted that scaling up the param-

 1. A summary of commonly used image encoders.

Variants	Pretraining Corpus	Resolution	Samples (B)	Parameter Size (M)
OpenCLIP-ConvNext-L [33]	LAION-2B	320	29	197.4
CLIP-ViT-L/14 [12]	OpenAI's WIT	224/336	13	304.0
EVA-CLIP-ViT-G/14 [34]	LAION-2B, COYO-700M	224	11	1000.0
OpenCLIP-ViT-G/14 [33]	LAION-2B	224	34	1012.7
OpenCLIP-ViT-bigG/14 [33]	LAION-2B	224	34	1844.9
InternViT-6B [35]	Multiple datasets	448	-	5540.0

 2. A summary of commonly used open-sourced LLMs. en, zh, fr, and de stand for English, Chinese, French, and German, respectively.

Model	Release Date	Pretrain Data Scale	Parameter Size (B)	Language Support	Architecture
Flan-T5-XL/XXL [44]	Oct-2022	-	3/11	en, fr, de	Encoder-Decoder
LLaMA [45]	Feb-2023	1.4T tokens	7/13/33/65	en	Causal Decoder
Vicuna [46]	Mar-2023	1.4T tokens	7/13/33	en	 Decoder
LLaMA-2 [47]	Jul-2023	2T tokens	7/13/70	en	 Causal Decoder
Qwen [48]	Sep-2023	3T tokens	1.8/7/14/72	en, zh	Causal Decoder
LLaMA-3 [49]	Apr-2024	15T tokens	8/70/405	en, fr, de, etc.	Causal Decoder

eter size of LLMs also brings additional gains, similar to the case of increasing input resolution. Specifically, Liu *et al.* [39,52] find that simply scaling up LLM from 7B to 13B brings comprehensive improvement on various benchmarks. Furthermore, when using a 34B LLM, the model shows emergent zero-shot Chinese capability, given that only English multimodal data are used during training. Li *et al.* [53] see a similar phenomenon by scaling up LLMs from 13B to 35B and 65B/70B, where the larger model size brings consistent gains on benchmarks specifically designed for MLLMs. We works instead use smaller LLMs to facilitate deployment on mobile devices. For example, MobileVLM series [54] use downscaled LLaMA [45] to enable efficient inference on mobile processors.

Recently, explorations of Mixture of Experts (MoE) architecture for LLMs have garnered rising attention [55]. Compared with dense models, the sparse architecture enables scaling up total parameter size without increasing computational cost, by selective activation of the parameters. Empirically, MM1 [41] and MoE-LLaVA [56] find that MoE implementation achieves better performance than the dense counterpart on almost all the benchmarks.

Modality interface

LLMs can only perceive text, bridging the gap between natural language and other modalities is necessary. Nevertheless, it would be costly to train from scratch a large multimodal model in an end-to-end manner. More practical way is to introduce a learnable connector between the pre-trained visual encoder and LLM. Another approach is to translate images into languages with the help of expert models, and then send

the language to LLM.

Learnable Connector. It is responsible for bridging the gap between different modalities. Specifically, the module projects information into the space that LLM can understand efficiently. Based on how multimodal information is fused, there are broadly two ways to implement such interfaces, *i.e.* token-level and feature-level fusion for different modalities.

Token-level fusion, features output from encoders are transformed into tokens and concatenated with text tokens before being sent into LLMs. A common solution is to leverage a group of learnable query tokens to extract information in a query-based manner [57], which first has been implemented in BLIP-2 [50], and subsequently inherited by a variety of work [22,51]. Q-Former-style approaches compress visual tokens into a smaller number of representation vectors. In contrast, some methods simply use a MLP-based interface to bridge the modality gap [16]. For example, LLaVA series adopts an MLP [16,39] to project visual tokens and align the feature dimension with word embeddings. LLaVA [58] adopts an ensemble of MLP-based and Q-Former-based connector to enhance performance in text-rich scenarios.

Another line, feature-level fusion inserts extra modules that enable deep interaction and fusion between text features and visual features. For example, Flamingo [59] inserts extra cross-attention layers between frozen Transformer layers of LLMs, thereby augmenting language features with external visual cues. Similarly, CogVLM [60] plugs in a visual expert module in each Transformer layer to enable dual interaction and fusion between vision and language features. For better performance, the QKV weight matrix of the introduced module

is initialized from the pre-trained LLM. **•**wise, LLaMA-Adapter [61] introduces learnable prompts into Transformer layers. **•**se prompts are first embedded with visual knowledge and then concatenated with text features as prefixes.

•a related note, MM1 [41] has ablated on design choices on the connector and found that for token-level fusion, the type of modality adapter is far less important than the number of visual tokens and input resolution. **•**ertheless, Zeng *et al.* [62] compare the performance of token and feature-level fusion, and empirically reveal that the token-level fusion variant performs better in terms of VQA benchmarks. **•**arding the performance gap, the authors suggest that cross-attention models might require a more complicated hyper-parameter searching process to achieve comparable performance.

•rms of parameter size, learnable interfaces generally comprise a small portion compared with encoders and LLMs. **•**e Qwen-VL [28] as an example, the parameter size of the Q-Former is about 0.08B, accounting for less than 1% of the whole parameters, while the encoder and the LLM account for about 19.8% (1.9B) and 80.2% (7.7B), respectively.

Expert Model. **•**t from the learnable interface, using expert models, **such as an image captioning model**, is also a feasible way to bridge the modality gap [63]. **•** basic idea is to convert multimodal inputs into languages without training. **•**is way, LLMs can understand multimodality by the converted languages. **•**example, VideoChat-Text [21] uses pre-trained vision models to extract visual information such as actions and enriches the descriptions using a speech recognition model. **•**ugh using expert models is straightforward, it may not be as flexible as adopting a learnable interface. **•**conversion of foreign modalities into text would cause information loss. **•** example, transforming videos into textual descriptions distorts spatial-temporal relationships [21].

TRAINING STRATEGY AND DATA

•ll-fledged MLLM undergoes three stages of training, *i.e.* pre-training, instruction-tuning, and alignment tuning. **•**ch phase of training requires different types of data and fulfills different objectives. **•**this section, we discuss training objectives, as well as data collection and characteristics for each training stage.

Input: <image>
Response: {caption}

•e 3. A simplified template to structure the caption data. <image> is the placeholder for the visual tokens, and {caption} is the caption for the image. Note that only the part marked in red is used for loss calculation.

Pre-training

Training detail

•he first training stage, pre-training mainly aims to align different modalities and learn multimodal world knowledge. **•**-training stage generally entails large-scale text-paired data, *e.g.* caption data. **•**ically, the caption pairs describe images/audio/videos in natural language.

•e, we consider a common scenario where MLLMs are trained to align vision with text. **•**llustrated in Table 3, given an image, the model is trained to predict autoregressively the caption of the image, following a standard cross-entropy loss. **•**mon approach for pre-training is to freeze pre-trained modules (*e.g.* visual encoders and LLMs) and train a learnable interface [16]. **•**idea is to align different modalities without losing pre-trained knowledge. **•**ne methods [28] also unfreeze more modules (*e.g.* visual encoder) to enable more trainable parameters for alignment. **•**ould be noted that the training scheme is closely related to data quality. **•**hort and noisy caption data, using lower resolution (*e.g.* 224) can speed up the training process, while for longer and cleaner data, it is better to utilize higher resolutions (*e.g.* 448 or higher) to mitigate hallucinations. **•**sides, ShareGPT4V [64] finds that with high-quality caption data in the pretraining stage, unlocking the vision encoder promotes better alignment.

•aining data mainly serve two purposes, *i.e.* (1) aligning different modalities and (2) providing world knowledge. **•**pretraining corpora can be divided into coarse-grained and fine-grained data according to granularities, which we will introduce sequentially. **•**summarize commonly used pretraining datasets in Table 4.

•se-grained caption data share some typical traits in common: **•**The data volume is large since samples are generally sourced from the internet. **•**Because of the web-scrawled nature, the captions are usually short and noisy since they originate from the alt-text of the web images. **•**se data can be cleaned and filtered via automatic tools, for example, using

CLIP [12] model to filter out image-text pairs whose similarities are lower than a pre-defined threshold. In what follows, we introduce some representative coarse-grained datasets.

CC-3M [65] is a web-scale caption dataset of 3.3M image-caption pairs, where the raw descriptions are derived from alt-text associated with images. The authors design a complicated pipeline to clean data: for images, those with inappropriate content or aspect ratio are filtered. For text, NLP tools are used to obtain text annotations, with samples filtered according to the designed heuristics. For image-text pairs, images are assigned labels via classifiers. If text annotations do not overlap with image labels, the corresponding samples are dropped.

CC-12M [66] is a following work of CC-3M and contains 12.4M image-caption pairs. Compared with the previous work, CC-12M relaxes and simplifies the data-collection pipeline, thus collecting more data.

SBU Captions [67]. A captioned photo dataset containing 1M image-text pairs, with images and descriptions sourced from Flickr. Specifically, an initial set of images is acquired by querying the Flickr website with a large number of query terms. Descriptions attached to the images thus serve as captions. In, to ensure that descriptions are relevant to the images, the retained images fulfill these requirements: Descriptions of the images are of satisfactory length, decided by observation. Options should contain at least 2 words in the predefined term lists and a propositional word (e.g. “on”, “under”) that suggests spatial relationships.

LAION. A series are large web-scale datasets, with images scrawled from the internet and associated alt-text as captions. After the image-text pairs, the following steps are performed: Text with short lengths or images with too small or too big sizes are dropped. Image deduplication based on URL. Extract CLIP [12] embeddings for images and text, and use the embeddings to drop possibly illegal content and image-text pairs with low cosine similarity between embeddings. We offer a brief summary of some typical variants:

- LAION-5B [68]: It is a research-purpose dataset of 5.85B image-text pairs. The dataset is multilingual with a 2B English subset.
- LAION-COCO [69]: It contains 600M images extracted from the English subset of LAION-5B. Captions are synthetic, using BLIP [70] to generate various image captions and using CLIP [12] to pick the best fit.

COYO-700M [71]. Contains 747M image-text

4. Common datasets used for pre-training.

Dataset	Samples	Date
Coarse-grained Image-Text		
CC-3M [65]	3.3M	2018
CC-12M [66]	12.4M	2020
SBU Captions [67]	1M	2011
LAION-5B [68]	5.9B	Mar-2022
LAION-2B [68]	2.3B	Mar-2022
LAION-COCO [69]	600M	Sep-2022
COYO-700M [71]	747M	Aug-2022
Fine-grained Image-Text		
ShareGPT4V-PT [64]	1.2M	Nov-2023
LVIS-Instruct4V [72]	111K	Nov-2023
ALLaVA [73]	709K	Feb-2024
Video-Text		
MSR-VTT [74]	200K	2016
Audio-Text		
WavCaps [75]	24K	Mar-2023

pairs, which are extracted from CommonCrawl. In terms of data filtering, the authors design the following strategies to filter out data samples: For images, those with inappropriate size, content, format, or aspect ratio are filtered. Moreover, the images are filtered based on the pHash value to remove images overlapped with public datasets such as ImageNet and MS-COCO. For text, only English text with satisfactory length, noun forms, and appropriate words are saved. Whitespace before and after the sentence will be removed, and consecutive whitespace characters will be replaced with a single whitespace. Moreover, text appearing more than 10 times (e.g. “image for”) will be dropped. For image-text pairs, duplicated samples are removed based on (image pHash, text) tuple.

Recently, more works [64,73] have explored generating high-quality fine-grained data through prompting strong MLLMs (e.g. GPT-4V). Compared with coarse-grained data, these data generally contain longer and more accurate descriptions of the images, thus enabling finer-grained alignment between image and text modalities. However, since the approach generally requires calling commercial-use MLLMs, the cost is higher, and the data volume is smaller. Notably, ShareGPT4V [64] strikes a balance by first training a captioner with GPT-4V-generated 100K data, then scaling up the data volume to 1.2M using the pre-trained captioner.

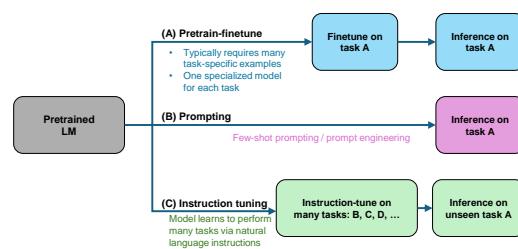


Fig. 3. Comparison of three typical learning paradigms, adapted from [76].

Instruction-tuning

Definition

Instruction refers to the description of tasks. Intuitively, instruction tuning aims to teach models to better understand the instructions from users and fulfill the demanded tasks. Using in this way, LLMs can generalize to unseen tasks by following new instructions, thus boosting zero-shot performance. This simple yet effective idea has sparked the success of subsequent NLP works, such as ChatGPT [77], InstructGPT [78].

Comparisons between instruction tuning and related typical learning paradigms are illustrated in Fig. 3. Supervised fine-tuning approach usually requires a large amount of task-specific data to train a task-specific model. Prompting approach reduces the reliance on large-scale data and can fulfill a specialized task via prompt engineering. Such a case, though the few-shot performance has been improved, the zero-shot performance is still quite average [4]. Interestingly, instruction tuning learns how to generalize to unseen tasks rather than fitting specific tasks like the two counterparts. Moreover, instruction tuning is highly related to multi-task prompting [79] and learning [80].

In this section, we delineate the format of instruction samples, the training objectives, typical ways to gather instruction data, and corresponding commonly used datasets.

Sampling detail

Multimodal instruction sample often includes an optional instruction and an input-output pair. Instruction is typically a natural language sentence describing the task, such as, “*Describe the image in detail.*” Input can be an image-text pair like the VQA task [82] or only an image like the image caption task [83]. Output is the answer to the instruction conditioned on the input. Instruction template is flexible and subject to manual designs [21], as exemplified in Table 5. Note that the instruction template can also be generalized to the case of multi-round human-agent conversations [16,81].

Finally, a multimodal instruction sample can be denoted in a triplet form, i.e. $(\mathcal{I}, \mathcal{M}, \mathcal{R})$, where $\mathcal{I}, \mathcal{M}, \mathcal{R}$ represent the instruction, the multimodal input, and the ground truth response, respectively. MLLM predicts an answer given the instruction and the multimodal input:

$$\mathcal{A} = f(\mathcal{I}, \mathcal{M}; \theta) \quad (1)$$

\mathcal{A} denotes the predicted answer, and θ are the parameters of the model. training objective is typically the original auto-regressive objective used to train LLMs [16], based on which the MLLM is encouraged to predict the next token of the response sequentially:

$$\mathcal{L}(\theta) = - \sum_{i=1}^N \log p(\mathcal{R}_i | \mathcal{I}, \mathcal{R}_{<i}; \theta) \quad (2)$$

where N is the length of the ground-truth.

Instruction collection

Instruction data are more flexible in formats and varied in task formulations, it is usually trickier and more costly to collect data samples. In this section, we summarize three typical ways to harvest instruction data at scale, i.e. data adaptation, self-instruction, and data mixture.

Adaptation. Task-specific datasets are rich sources of high-quality data. Previous, abundant works [51,84] have utilized existing high-quality datasets to construct instruction-formatted datasets. Take the transformation of VQA datasets for an example: the original sample is an input-out pair where the input comprises an image and a natural language question, and the output is the textual answer to the question conditioned on the image. Input-output pairs of these datasets could naturally comprise the multimodal input and response of the instruction sample. Instructions, i.e. the descriptions of the tasks, can either derive from manual

Instruction: *<instruction>*
Input: {*<image>, <text>*}
Response: *<output>*

Table 5. Simplified template to structure the multimodal instruction data. *<instruction>* is a textual description of the task. *<image>, <text>* and *<output>* are input and output from the data sample. Note that *<text>* in the input may be missed for some datasets, such as image caption datasets merely have *<image>*. This example is adapted from [81].

- <Image> {Question}
- <Image> Question: {Question}
- <Image> {Question} A short answer to the question is
- <Image> Q: {Question} A:
- <Image> Question: {Question} Short answer:
- <Image> Given the image, answer the following question with no more than three words. {Question}
- <Image> Based on the image, respond to this question with a short answer: {Question}. Answer:
- <Image> Use the provided image to answer the question: {Question} Provide your answer as short as possible:
- <Image> What is the answer to the following question? "{Question}"
- <Image> The question "{Question}" can be answered using the image. A short answer is

Table 6. Instruction templates for VQA datasets, cited from [51]. <Image> and {Question} are the image and the question in the original VQA datasets, respectively.

design or from semi-automatic generation aided by GPT. Specifically, some works [17] handcraft a pool of candidate instructions and sample one of them during training. [51] offer an example of instruction templates for the VQA datasets as shown in Table 6. [51] other works manually design some seed instructions and use these to prompt GPT to generate more [21].

[51] Note that since the answers of existing VQA and caption datasets are usually concise, directly using these datasets for instruction tuning may limit the output length of MLLMs. [51] There are two common strategies to tackle this problem. [51] The first one is to specify the corresponding requirements explicitly in the instructions. [51] For example, ChatBridge [85] explicitly declares *short* and *brief* for short-answer data. [51] The second one is to extend the length of existing answers [86]. [51] For example, M³IT [86] proposes to rephrase the original answer by prompting ChatGPT with the original question, answer, and contextual information of the image (*e.g.* caption and text extracted through OCR).

Instruction. [51] Though existing multi-task datasets can contribute a rich source of data, they usually do not meet human needs well in real-world scenarios, such as multiple-round conversations. [51] To tackle this issue, some works collect samples through self-instruction [89], which utilizes LLMs to generate textual instruction-following data using a few hand-annotated samples. [51] Specifically, some instruction-following samples are hand-crafted as demonstrations, after which ChatGPT/GPT-4 is prompted to generate more instruction samples with the demonstrations as guidance. [51] iVA [16] extends the approach to the multimodal field by translating images into text of captions and bounding boxes, and prompting text-only GPT-4 to gen-

erate new data with the guidance of requirements and demonstrations. [51] This way, a multimodal instruction dataset is constructed, called **LLaVA-Instruct-150k**. [51] Owing this idea, subsequent works such as MiniGPT-4 [17] and GPT4Tools [90] develop different datasets catering to different needs. [51] Recently, with the release of the more powerful multimodal model GPT-4V, many works have adopted GPT-4V to generate data of higher quality, as exemplified by LVIS-Instruct4V [72] and ALLaVA [73]. [51] To summarize the popular datasets generated through self-instruction in Table 7. [51] It should be noted that this paradigm highly relies on advanced yet close-sourced models, which can be expensive for data scaling. [51] Its approach might be partially due to the limited capabilities of early models. [51] Future research can explore capitalizing on open-sourced models to generate high-quality instruction data.

Mixture. [51] Apart from the multimodal instruction data, language-only user-assistant conversation data can also be used to improve conversational proficiencies and instruction-following abilities [91]. [51] IN [91] directly constructs a minibatch by randomly sampling from both language-only and multimodal data. [51] iInstruct [84] probes different strategies for training with a fusion of single modal and multimodal data, including mixed instruction tuning (combine both types of data and randomly shuffle) and sequential instruction tuning (text data followed by multimodal data).

quality

[51] Recent research has revealed that the data quality of instruction-tuning samples is no less important than quantity. [51] x [62] finds that models pre-trained on large-scale but noisy image-

Table 7. A summary of popular datasets generated by self-instruction. I: Image, T: Text, V: Video, A: Audio. Data composition, M-T and S-T denote multi-turn and single-turn, respectively.

Dataset	Sample	Modality	Source	Composition
LLaVA-Instruct [16]	158K	I + T → T	MS-COCO	23K caption + 58K M-T QA + 77K reasoning
LVIS-Instruct [72]	220K	I + T → T	LVIS	110K caption + 110K M-T QA
ALLaVA [73]	1.4M	I + T → T	VFlan, LAION	709K caption + 709K S-T QA
Video-ChatGPT [87]	100K	V + T → T	ActivityNet	7K description + 4K M-T QA
VideoChat [21]	11K	V+T → T	WebVid	description + summarization + creation
Clotho-Detail [88]	3.9K	A + T → T	Clotho	caption

text pairs do not perform as well as models pre-trained with smaller but cleaner datasets. Similarly, Wei *et al.* [92] finds that less instruction-tuning data with higher quality can achieve better performance. In data filtering, the work proposes some metrics to evaluate data quality and, correspondingly, a method to automatically filter out inferior vision-language data. Here we discuss two important aspects of data quality.

Inpt Diversity. The diversity of instructions has been found to be critical for model performance. Xu [62] empirically verifies that diverse prompts help improve model performance and generalization ability.

Coverage. In terms of tasks involved in training data, Du *et al.* [93] perform an empirical study and find that the visual reasoning task is superior to captioning and QA tasks for boosting model performance. Moreover, the study suggests that more complex instructions are better than increasing task diversity and incorporating fine-grained spatial annotations.

Alignment tuning

Introduction

Alignment tuning is more often used in scenarios where models need to be aligned with specific human preferences, *e.g.* response with fewer hallucinations. Recently, Reinforcement Learning with Human Feedback (RLHF) and Direct Preference Optimization (DPO) are two main techniques for alignment tuning. In this section, we introduce the main ideas of the two techniques in sequence and offer some examples of how they are utilized in addressing practical problems, and finally, give a compilation of the related datasets.

Training detail

RLHF [94,95]. This technique aims to utilize reinforcement learning algorithms to align LLMs with human preferences, with human annotations as supervision in the training loop. exemplified in InstructGPT [78], RLHF incorporates three key steps:

(1) **Supervised fine-tuning.** This step aims to fine-tune a pre-trained model to present the preliminary desired output behavior. The fine-tuned model in the RLHF setting is called a *policy model*. Note that this step might be skipped since the supervised policy model π^{SFT} can be initialized from an instruction-tuned model.

Reward modeling. The reward model is trained using preference pairs in this step. Given a multimodal prompt (*e.g.* image and text) x and a response pair (y_w, y_l) , the reward model r_θ learns to give a higher reward to the preferred response y_w , and vice versa for y_l , with the objective:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log(\sigma(r_\theta(x, y_w)) - r_\theta(x, y_l)) \right] \quad (3)$$

The $\mathcal{D} = \{(x, y_w, y_l)\}$ is the comparison dataset labeled by human annotators. In practice, the reward model r_θ shares a similar structure with the policy model.

Reinforcement learning. In this step, the Proximal Policy Optimization (PPO) algorithm is adopted to optimize the RL policy model π_ϕ^{RL} . Per-token KL penalty is often added to the training objective to avoid deviating too far from the original policy [78], resulting in the objective:

$$\mathcal{L}(\phi) = -\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\phi^{\text{RL}}(y|x)} \left[r_\theta(x, y) - \beta \cdot \mathbb{D}_{KL}(\pi_\phi^{\text{RL}}(y|x) || \pi^{\text{REF}}(y|x)) \right] \quad (4)$$

where β is the coefficient for the KL penalty term. Typically, both the RL policy π_ϕ^{RL} and the reference model π^{REF} are initialized from the supervised model π^{SFT} . The obtained RL policy model is expected to align with human preferences through this tuning process.

Researchers have explored using the RLHF techniques for better multimodal alignment.

 8. A summary of datasets for alignment-tuning.
Input/output modalities, I: Image, T: Text.

Dataset	Sample	Modality	Source
LLaVA-RLHF [96]	10K	I + T → T	Human
RLHF-V [98]	5.7K	I + T → T	Human
VLFeedback [99]	380K	I + T → T	GPT-4V

example, LLaVA-RLHF [96] collects human preference data and tunes a model with fewer hallucinations based on LLaVA [16].

DPO [97].  learns from human preference labels utilizing a simple binary classification loss.  Compared with the PPO-based RLHF algorithm, DPO is exempt from learning an explicit reward model, thus simplifying the whole pipeline to two steps, *i.e.* human preference data collection and preference learning.  learning objective for the algorithm is as follows:

$$\mathcal{L}(\phi) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\phi}^{\text{RL}}(y_w|x)}{\pi^{\text{REF}}(y_w|x)} \right) \right. \\ \left. - \beta \log \frac{\pi_{\phi}^{\text{RL}}(y_l|x)}{\pi^{\text{REF}}(y_l|x)} \right] \quad (5)$$

 IF-V [98] collects fine-grained (segment-level) preference data pairs by correcting hallucinations in the model response and uses the obtained data to perform dense DPO.  ie [99] instead collects preference data via prompting GPT-4V and distills the preference supervision into an instruction-tuned model through DPO.

 list of data collection for alignment-tuning is to collect feedback for model responses, *i.e.* to decide which response is better.  generally more expensive to collect such data, and the amount of data used for this phase is typically even less than that used in previous stages.  this part, we introduce some datasets and summarize them in Table 8.

LLaVA-RLHF [96].  ntains 10K preference pairs collected from human feedback in terms of honesty and helpfulness.  dataset mainly serves to reduce hallucinations.

RLHF-V [98].  as 5.7K fine-grained human feedback data collected by performing segment-level hallucination corrections.

VLFeedback [99].  ilizes AI to provide feedback on model responses.  dataset contains more than 380K comparison pairs scored by GPT-4V in terms of helpfulness, faithfulness, and ethical concerns.

9. Evaluation

Evaluation is an essential part of developing MLLMs since it provides feedback for model optimization and helps to compare the performance of different models. pared with evaluation methods of traditional multimodal models, the evaluation of MLLMs exhibits several new traits:  Since MLLMs are generally versatile, it is important to evaluate MLLMs comprehensively.  MLLMs exhibit many emergent capabilities that require special attention (*e.g.* OCR-free math reasoning) and thus require new evaluation schemes.  evaluation of MLLMs can be broadly categorized into two types according to the question genres, including closed-set and open-set.  sed-set evaluation often involves task-specific benchmarks and more comprehensive benchmarks specifically designed for MLLM, where answers are limited to predefined sets.  h-set evaluation typically includes manual scoring, GPT scoring, and case study.

Closed-set

Closed-set questions refer to a type of question where the possible answer options are predefined and limited to a finite set.  evaluation is usually performed on task-specific datasets.  his case, the responses can be naturally judged by benchmark metrics.  example, InstructBLIP [51] reports the accuracy on ScienceQA [100], as well as the CIDEr score [101] on NoCaps [102].  evaluation settings are typically zero-shot [51,84] or finetuning [29,51].  first setting often selects a wide range of datasets covering different general tasks and splits them into held-in and held-out datasets.  tuning on the former, zero-shot performance is evaluated on the latter with unseen datasets or even unseen tasks. ontrast, the second setting is often observed in the evaluation of domain-specific tasks.  example, LLaVA [16] reports finetuned performance on ScienceQA [100]. VA-Med [29] reports results on biomedical VQA [103].

The above evaluation methods are usually limited to a small range of selected tasks or datasets, lacking a comprehensive quantitative comparison. his end, some efforts have endeavored to develop new benchmarks specially designed for MLLMs [104,105].  example, Fu *et al.* [104] construct a comprehensive evaluation benchmark MME that includes a total of 14 perception and cognition tasks.  instruction-answer pairs in MME are manually

designed to avoid data leakage. Bench [105] is a benchmark specifically designed for evaluating multiple dimensions of model capabilities, using ChatGPT to match open responses with pre-defined choices. Vleo-ChatGPT [87] and Video-Bench [106] focus on video domains and propose specialized benchmarks as well as evaluation tools for assessment.

Open-set

In contrast to the closed-set questions, the responses to open-set questions can be more flexible, where MLLMs usually play a chatbot role. Since the content of the chat can be arbitrary, it would be trickier to judge than the closed-ended output. This criterion can be classified into manual scoring, GPT scoring, and case study. Manual scoring requires humans to assess the generated responses. This kind of approach often involves hand-crafted questions that are designed to assess specific dimensions. For example, mPLUG-Owl [107] collects a visually related evaluation set to judge capabilities like natural image, diagram, and flowchart understanding. Similarly, GPT4Tools [90] builds two sets for the finetuning and zero-shot performance, respectively, and evaluates the responses in terms of thought, action, arguments, and the whole.

The manual assessment is labor intensive, some researchers have explored rating with GPT, namely GPT scoring. This approach is often used to evaluate performance on multimodal dialogue. AVA [16] proposes to score the responses via text-only GPT-4 in terms of different aspects, such as helpfulness and accuracy. Specifically, 30 images are sampled from the COCO [108] validation set, each associated with a short question, a detailed question, and a complex reasoning question via self-instruction on GPT-4. The answers generated by both the model and GPT-4 are sent to GPT-4 for comparison. Subsequent works follow this idea and prompt ChatGPT or GPT-4 to rate results [29] or judge which one is better [109].

A main issue of applying text-only GPT-4 for evaluation is the judge is only based on translated text content, such as captions or bounding box coordinates, without accessing the image [29]. Thus, it may be questionable to set GPT-4 as the performance upper bound in this case. With the release of the vision interface of GPT, some works exploit the more advanced GPT-4V model to assess the performance of MLLMs. For example, Woodpecker [63] adopts the GPT-4V model to judge the response quality of model answers. Evaluation is expected

to be more accurate than using text-only GPT-4 since GPT-4V has direct access to the image.

While the benchmark evaluation is not comprehensive enough, a supplementary approach is to compare the different capabilities of MLLMs through case studies. For instance, some studies evaluate two typical advanced commercial-use models, GPT-4V and Gemini. Li et al. [110] perform in-depth qualitative analysis on GPT-4V by crafting a series of samples across various domains and tasks, spanning from preliminary skills, such as caption and object counting, to complex tasks that require world knowledge and reasoning, such as joke understanding and indoor navigation as an embodied agent. Lin et al. [111] make a more focused evaluation of GPT-4V by designing samples targeting automatic driving scenarios. Li et al. [112] carry out a comprehensive evaluation on Gemini-Pro by comparing the model against GPT-4V. The results suggest that GPT-4V and Gemini exhibit comparable visual reasoning abilities in spite of different response styles.

EXENSIONS

Recent studies have made significant strides in extending the capabilities of MLLMs, spanning from more potent foundational abilities to broader coverage of scenarios. We trace the principal development of MLLMs in this regard.

Granularity Support. To facilitate better interaction between agents and users, researchers have developed MLLMs with finer support of granularities in terms of model inputs and outputs. On the input side, models that support finer control from user prompts are developed progressively, evolving from image to region [24] and even pixels [25]. Specifically, Shikra [24] supports region-level input and understanding. This may interact with the assistant more flexibly by referring to specific regions, which are represented in bounding boxes of natural language forms. Ret [113] takes a step further and supports more flexible referring by devising a hybrid representation scheme. The model supports different forms of prompts, including point, box, and sketch. Similarly, Osprey [25] supports point input by utilizing a segmentation model [10]. Led by the exceptional capabilities of the pre-trained segmentation model, Osprey enables specifying a single entity or part of it with a single click. On the output side, **grounding capabilities** are improved in line with the development of input support. Shikra [24] supports response grounded in the image with box annotations, resulting in higher precision and finer re-

ferring experience. A [114] further supports mask-level understanding and reasoning, which makes pixel-level grounding possible.

Modality Support. Increased support for modalities is a tendency for MLLM studies. On the one hand, researchers have explored adapting MLLMs to support the input of more multimodal content, such as 3D point cloud [115]. On the other hand, MLLMs are also extended to generate responses of more modalities, such as image [116], audio [117], and video [118]. For example, NExT-GPT [119] proposes a framework that supports inputs and outputs of mixed modalities, specifically, combinations of text, image, audio, and video, with the help of diffusion models [120] attached to the MLLM. This framework applies an encoder-decoder architecture and puts LLM as a pivot for understanding and reasoning.

Language Support. Current models are predominantly unilingual, probably due to the fact that high-quality non-English training corpus is scarce. Some works have been devoted to developing multilingual models so that a broader range of users can be covered. CPM [121] transfers model capabilities to the multilingual setting by designing a multi-stage training scheme. Specifically, the scheme takes English as a pivotal language, with abundant training corpus. Utilizing a pre-trained bilingual LLM, the multimodal capabilities are transferred to Chinese by adding some translated samples during instruction tuning. Using a similar approach, Qwen-VL [28] is developed from the bilingual LLM Qwen [48] and supports both Chinese and English. During pre-training, Chinese data is mixed into the training corpus to preserve the bilingual capabilities of the model, taking up 22.7% of the whole data volume.

Scenario/Task Extension. Apart from developing common general-purpose assistants, some studies have focused on more specific scenarios where practical conditions should be considered, while others extend MLLMs to downstream tasks with specific expertise.

Typical tendency is to adapt MLLMs to more specific real-life scenarios. For example, some works develop agents that interact with the real world, *e.g.* user-friendly assistants specially designed for Graphical User Interface (GUI), as exemplified by CogAgent [32], AppAgent [122], and Mobile-Agent [123]. Researchers also develop embodied agents [19,31] that can perform reasoning, navigation, and manipulation in the real world, facilitating the development of automatic agents that can execute tasks for humans. Generally, these assistants excel in planning and

performing each step to fulfill tasks specified by users, acting as helpful agents for humans.

Another line is to augment MLLMs with specific skills for solving tasks in different domains, *e.g.* document understanding [30] and medical domains [29]. For document understanding, mPLUG-DocOwl [124] utilizes various forms of document-level data for tuning, resulting in an enhanced model in OCR-free document understanding. Monkey [30] incorporates multiple tasks related to document understanding to improve model performance. Similarly, MLLMs can also be trained to accommodate traditional vision tasks such as visual grounding [125,126]. Compared with traditional methods [13,127], MLLMs unify the I/O format and streamline the whole learning and inference process. Specifically, it is feasible to recast the grounding task into a conditioned box coordinate prediction task under a unified language modeling objective [24,28,52]. The model is trained to predict the coordinates of specified objects in the form of natural language. MLLMs can also be extended to medical domains by instilling specialized knowledge. For example, LLaVA-Med [29] develops assistants specialized in medical image understanding and question answering by injecting domain knowledge.

Efficient MLLMs. Recently, using lightweight MLLMs for efficient deployment has gained increased popularity [128–130]. These models are meticulously designed and optimized for more economical utilization or resource-limited scenarios without compromising too much on model performance.

From a model perspective, various techniques have been explored to facilitate efficient training and inference. For instance, MobileVLM [54] explores developing small-size variants of MLLMs for resource-limited scenarios. The designs and techniques are utilized for deployment on mobile devices, such as LLMs of smaller size and quantization techniques to speed up computation. Similarly, MiniCPM-V [129] builds efficient MLLMs for end-side computation. ViFormer [28] is adopted to cut down the number of visual tokens for each patch of the image.

From a data perspective, Bunny [130] comprehensively investigates efficient data selection and combination schemes for model training. The obtained models achieve performance on par with MLLMs of larger parameter sizes.

TIMODAL HALLUCINATION

 imodal hallucination refers to the phenomenon of responses generated by MLLMs being inconsistent with the image content [63].  fundamental problem has received increased attention.  his section, we briefly introduce related concepts and research development.

Preliminaries

 imodal hallucinations can be categorized into three types [131]:

 *Existence Hallucination* is a common type, meaning that models incorrectly decide the existence of objects.

 *Attribute Hallucination* means falsely describing the attributes of certain objects, *e.g.* failure to identify a dog's color.

 *Relationship Hallucination* is a more complex type of hallucination.  refers to false descriptions of relationships between objects, such as relative positions.

 hat follows, we first introduce evaluation methods, which are useful to gauge the performance of methods for mitigating hallucinations.

 n, we discuss mitigation methods of different kinds of approaches.

Evaluation methods

 IR [132] is an early metric that evaluates hallucination levels in open-ended captions.  metric measures the proportion of sentences with hallucinated objects or the proportion of hallucinated objects in all the objects mentioned.  ontrast, POPE [133] is a method that evaluates closed-set choices.  Specifically, multiple prompts with binary choices are formulated, each querying if a specific object exists in the image.  h a similar evaluation approach, MME [104] provides a more comprehensive evaluation, covering aspects of existence, count, position, and color, as exemplified in [63].

 erent from previous approaches that use matching mechanisms to detect and decide hallucinations, some works explore automatic evaluation of text responses via models.  example, HaELM [134] proposes using LLMs as a judge to decide whether MLLMs' captions are correct against reference captions.  view of the fact that text-only LLMs can only access limited image context and require reference annotations, Woodpecker [63] uses GPT-4V to directly assess model responses grounded in the image.

Mitigation methods

 ording to high-level ideas for mitigating hallucinations, current methods can be roughly divided into three categories: pre-correction, in-process-correction, and post-correction.

 **Pre-correction.**  intuitive solution for hallucination is to collect specialized data (*e.g.* negative data) and use the data for fine-tuning, thus achieving models with fewer hallucinations.

 -Instruction [135] introduces a visual instruction tuning dataset to encourage faithful generation.  iliarly, LLaVA-RLHF [96] collects human-preference pairs and finetunes models with reinforcement learning techniques.

 **In-process-correction.**  other line is to make improvements in architectural design or feature representation.  ese works try to explore the reasons for hallucinations and design remedies to mitigate them in the generation process.  example, Halle-Switch [131] introduces a continuous controlling factor to control the extent of imagination in model output during inference.

 **Post-correction.**  ferent from previous paradigms, post-correction mitigates hallucinations in a post-remedy way.  example, Woodpecker [63] is a training-free framework for hallucination correction.  cifically, the method incorporates expert models to supplement contextual information of the image and crafts a pipeline to correct hallucinations step by step.

ENDED TECHNIQUES

Unimodal in-context learning

 is one of the important **emergent abilities** of LLMs.  essence of the technique is prompting the model with a few examples as guidance to make it easier for the model to answer the query.  re are two good traits of ICL:  The crux of ICL is to learn from analogy [136], thus largely reducing the requirement of data samples.  CL is usually implemented in a training-free way [136] and can be flexibly integrated into various frameworks at inference time.

 he context of MLLM, ICL has been extended to more modalities, leading to Multi-modal ICL (M-ICL).  inference time, M-ICL can be implemented by adding a demonstration set, *i.e.* a set of in-context samples, to the original sample.  his case, the template can be extended as illustrated in Table 9.

Improvement on ICL capabilities

 ently, a growing amount of work has focused on enhancing ICL performance under various

<BOS> Now are some examples and an instruction that describes a task. Write a response that appropriately completes the request

Instruction: {instruction}

Image: <image>

Response: {response}

Image: <image>

Response: {response}

Image: <image>

Response: <EOS>

Table 9. A simplified example of the template to structure an M-ICL query, adapted from [81]. Illustration, we list two in-context examples and a query divided by a dashed line. {instruction} and {response} are texts from the data sample. <image> is a placeholder to represent the multimodal input (an image in this case). <S> and <EOS> are tokens denoting the start and the end of the input to the LLM, respectively.

scenarios. In this section, we trace the development of this field and summarize relevant works.

MIC-IT [137] combines in-context learning with instruction tuning by building an instruction dataset formatted with multimodal context. The other works explore improving few-shot learning performance under specific settings. For example, Link-context learning [138] focuses on the causal relationships between demonstrations and queries, and casts a contrast training scheme by formulating positive and negative image-description pairs. Similarly, Yang *et al.* [139,140] explore different strategies to optimize demonstration configurations (selections or orderings of in-context samples) to achieve better few-shot performance.

Applications

In terms of applications in multimodality, M-ICL is mainly used in two scenarios: (1) solving various visual reasoning tasks [141] and (2) teaching LLMs to use external tools [142,143]. The former involves learning from several task-specific examples and generalizing to a new but similar question. In contrast, examples of tool usage are more fine-grained, typically comprising a chain of steps to fulfill the task.

Multimodal chain of thought

It is “a series of intermediate reasoning steps” [5]. This technique has been proven to be effective in complex reasoning tasks.

main idea is to prompt LLMs to output not only the final answer but also the reasoning process that leads to the answer, resembling the cognitive process of humans.

Inspired by the success in NLP realms, multiple works [144,145] have proposed to extend the technique to Multimodal CoT (M-CoT). They first introduce different paradigms for acquiring the M-CoT ability. Then, we delineate more specific aspects of M-CoT, including the chain configuration and the pattern.

Learning paradigms

There are broadly three ways to acquire the M-CoT ability, *i.e.* through finetuning and training-free few/zero-shot learning.

Generally, the finetuning approach often involves curating specific datasets for M-CoT learning. For example, Lu *et al.* [100] construct a scientific question-answering dataset ScienceQA with lectures and explanations, which can serve as sources of learning CoT reasoning.

Compared with finetuning, few/zero-shot learning is more computationally efficient. Few-shot learning approach typically requires hand-crafted in-context examples to teach reasoning step by step. In contrast, the zero-shot learning approach directly prompts with designed instructions [144].

Chain configuration

Structure and length are two critical aspects of the reasoning chains. In terms of structure, current methods can be divided into single-chain [100] and tree-shape methods [146]. In length can be categorized into adaptive and pre-defined formations. The former configuration requires LLMs to decide when to halt the reasoning chains [100], while the latter setting stops the chains with a pre-defined length [147].

Generation patterns

We summarize the relevant works into (1) an infilling-based pattern and (2) a predicting-based pattern. Specifically, the infilling-based pattern demands deducing steps between surrounding context (previous and following steps) to fill the logical gaps [144]. In contrast, the predicting-based pattern requires extending the reasoning chains given conditions such as instructions and previous reasoning history [142].

Tool-Aided visual reasoning

Introduction

Inspired by the success of tool-augmented LLMs [148], some researchers have explored

the possibilities of invoking external tools or vision foundation models for visual reasoning tasks. Using LLMs as helpers with different roles, these works build task-specific or general-purpose visual reasoning systems.

Compared with conventional visual reasoning models, these works manifest several good traits: Strong generalization abilities. Equipped with rich open-world knowledge learned from large-scale pretraining, these systems can easily generalize to unseen objects or concepts with remarkable zero/few-shot performance [149]. Emergent abilities. Led by the strong reasoning abilities of LLMs, these systems can perform complex tasks, *e.g.* understanding the deep meaning of an image [18]. Better interactivity and control. Traditional models typically allow a limited set of control mechanisms, while LLM-based systems enable finer control in a user-friendly interface (*e.g.* click and natural language queries) [150].

In this part, we start with introducing different training paradigms employed in the construction of LLM-Aided Visual Reasoning systems. Then, we delve into the primary roles that LLMs play within these systems.

Training paradigms

According to training paradigms, LLM-Aided Visual Reasoning systems can be divided into two types, *i.e.* training-free and finetuning.

Training-free. With abundant prior knowledge stored in pre-trained LLMs, an intuitive and simple way is to freeze pre-trained models and directly prompt LLMs to fulfill various needs. According to the setting, the reasoning systems can be further categorized into few-shot models [142] and zero-shot models [150].

Finetuning. The works adopt further finetuning to improve the planning abilities with respect to tool usage [90] or to improve localization capabilities [114] of the system. For example, GPT4Tools [90] collects a tool-related instruction dataset to finetune the model.

Functions

Regarding what roles LLMs exactly play in LLM-Aided Visual Reasoning systems, existing related works are divided into three types:

- LLM as a Controller
- LLM as a Decision Maker
- LLM as a Semantics Refiner

Delineate how LLMs serve these roles in the following parts.

LLM as a Controller. In this case, LLMs act as a central controller that (1) breaks down a com-

plex task into simpler sub-tasks/steps and (2) assigns these tasks to appropriate tools/modules.

Specifically, LLMs are prompted explicitly to output task planning [151] or, more directly, the modules to call [90,142,143]. For example, Vis-Prog [143] prompts GPT-3 to output a visual program, where each program line invokes a module to perform a sub-task.

LLM as a Decision Maker. In this case, complex tasks are solved in a multi-round manner, often in an iterative way [152]. Vision-makers often (1) summarize the context to decide whether to finish the task and (2) organize the answer in a user-friendly way.

LLM as a Semantics Refiner. When LLM is used as a Semantics Refiner, researchers mainly utilize its rich linguistic and semantic knowledge. Specifically, LLMs are often instructed to integrate information into fluent natural language sentences [153] or generate texts according to different specific needs [149,150,154].

CHALLENGES AND FUTURE DIRECTIONS

The development of MLLMs is still in a rudimentary stage and thus leaves much room for improvement, which we summarize below:

- Current MLLMs are limited in processing multimodal information of long context. This restricts the development of advanced models with more multimodal tokens, *e.g.* long-video understanding and long documents interleaved with images and text.

LLMs should be upgraded to follow more complicated instructions. For example, a mainstream approach to generating high-quality question-answer pair data is still prompting closed-source GPT-4V because of its advanced instruction-following capabilities, while other models generally fail to achieve such goals.

There is still a large space for improvement in techniques like M-ICL and M-CoT. Current research on the two techniques is still rudimentary, and the related capabilities of MLLMs are still weak. Therefore, explorations on the underlying mechanisms and potential improvements are promising.

Developing embodied agents based on MLLMs is a heated topic. It would be meaningful to develop such agents that can interact with the real world. These endeavors require models with critical capabilities, including perception, reasoning, planning, and execution.

- Safety issues. Similar to LLMs, MLLMs can be vulnerable to crafted attacks. In other words, MLLMs can be misled to output biased or undesirable responses. Thus, improving model safety will be an important research topic.
- Interdisciplinary research. Given the strong generalization capabilities and abundant pre-trained knowledge of MLLMs, a promising research direction could be utilizing MLLMs to boost research fields of natural sciences, e.g. leveraging MLLMs for analysis of medical images or remote sensing images. To achieve this goal, injecting domain-specific multimodal knowledge into MLLMs might be necessary.

CONCLUSION

In this paper, we perform a survey of the existing MLLM literature and offer a broad view of its main directions, including the basic recipe and related extensions. Moreover, we underscore the current research gaps that need to be filled and point out some promising research directions. We hope this survey can offer readers a clear picture of the current progress of MLLM and inspire more relevant works. Given the fact that the era of MLLM has only just begun, we will keep updating this survey and hope it can inspire more research. An associated GitHub link collecting the latest papers is available [here](#).

FUNDING

This work was supported in part by the National Natural Science Foundation of China (62222213, 62406264, U22B2059, U23A20319, 62072423 and 61727809) and the Young Scientists Fund of the Natural Science Foundation of Sichuan Province (2023NSFSC1402).

AUTHOR CONTRIBUTIONS

C.F. was the project leader. S.Y., C.F. and S.Z. performed the literature review. K.L., X.S., T.X. and E.C. provided insights of the relevant fields. S.Y., C.F. and S.Z. wrote the article. All authors discussed and advised on the manuscript.

Conflict of interest statement. None declared.

REFERENCES

- Zhao WX, Zhou K, Li J et al. A survey of large language models. *arXiv:2303.18223*; . 1
- Xu B and Poo Mm. Large language models and brain-inspired general intelligence. *Natl Sci Rev* 2023; **10**: nwad267. 1
- Peng B, Li C, He P et al. Action tuning with gpt-4. *arXiv:2304.03277*; . 1
- Brown T, Mann B, Ryder N et al. Language models are few-shot learners. *Conference on Neural Information Processing Systems*, volume 33 (2020) 1877–1901. 1, 3, 7
- Wei J, Wang X, Schuurmans D et al. Few-of-thought prompting elicits reasoning in large language models. *Conference on Neural Information Processing Systems*, volume 35 (2022) 24824–24837. 1, 14
- Li H. Learning for natural language processing: advantages and challenges. *Natl Sci Rev* 2018; **5**: 24–26. 1
- Zhao W. Model discussion on ai for science: the opportunities, challenges and reflections. *Natl Sci Rev* 2024; nwae119. 1
- Xie WJ and Warshel A. Harnessing generative ai to decode enzyme catalysis and evolution for enhanced engineering. *Natl Sci Rev* 2023; **10**: nwad331. 1
- Gong P, Guo H, Chen B et al. An interdisciplinary framework in the era of big data and ai for sustainable development. *Natl Sci Rev* 2023; **10**: nwad178. 1
- Kirillov A, Mintun E, Ravi N et al. Agent anything. *IEEE/CVF International Conference on Computer Vision* (2023) 4015–4026. 1, 11
- Shen Y, Fu C, Chen P et al. Aligning and prompting everything all at once for universal visual perception. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 13193–13203. 1
- Radford A, Kim JW, Hallacy C et al. Aligning transferable visual models from natural language supervision. *International Conference on Machine Learning* (2021) 8748–8763. 1, 3, 4, 6
- Li J, Selvaraju R, Gotmare A et al. Align before fuse: Vision and language representation learning with momentum distillation. *Conference on Neural Information Processing Systems*, volume 34 (2021) 9694–9705. 1, 12
- Wang P, Yang A, Men R et al. Ofa: Aligning architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *International Conference on Machine Learning*, volume 162 (2022) 23318–23340. 1
- Cho J, Lei J, Tan H et al. Aligning vision-and-language tasks via text generation. *International Conference on Machine Learning* (2021) 1931–1942. 1
- Liu H, Li C, Wu Q et al. Multi instruction tuning. *Conference on Neural Information Processing Systems*, volume 36 (2024) . 1, 2, 4, 5, 7, 8, 9, 10, 11
- Zhu D, Chen J, Shen X et al. Gpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*; . 1, 3, 8
- Yang Z, Li L, Wang J et al. Chatgpt: Prompting chatgpt for multimodal reasoning and action. *arXiv:2303.11381*; . 1, 15
- Driess D, Xia F, Sajjadi MS et al. Gaze: An embodied multimodal language model. *International Conference on Machine Learning*, volume 202 (2023) 8469–8488. 1, 12

20. OpenAI. Technical report. *arXiv:2303.08774* ; . 1
21. Li K, He Y, Wang Y et al. Chat: Chat-centric video understanding. *arXiv:2305.06355* ; . 2, 5, 7, 8, 9
22. Zhang H, Li X and Bing L. -llama: An instruction-tuned audio-visual language model for video understanding. *Conference on Empirical Methods in Natural Language Processing* (2023) . 2, 4
23. Deshmukh S, Elizalde B, Singh R et al. Pengi: audio language model for audio tasks. *Conference on Neural Information Processing Systems 2023*; 36: 18090–18108. 2, 3
24. Chen K, Zhang Z, Zeng W et al. : Unleashing multi-modal llm's referential dialogue magic. *arXiv:2306.15195* ; . 2, 11, 12
25. Yuan Y, Li W, Liu J et al. : Pixel understanding with visual instruction tuning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 28202–28211. 2, 3, 11
26. Han J, Zhang R, Shao W et al. bind-llm: Multi-modality instruction tuning. *arXiv:2309.03905* ; . 2, 3
27. Moon S, Madotto A, Lin Z et al. al: An efficient and scalable any-modality augmented language model. *arXiv:2309.16058* ; . 2
28. Bai J, Bai S, Yang S et al. A frontier large vision-language model with versatile abilities. *arXiv:2308.12966* ; . 2, 3, 5, 12
29. Li C, Wong C, Zhang S et al. med: Training a large language-and-vision assistant for biomedicine in one day. *Conference on Neural Information Processing Systems*, volume 36 (2024) . 2, 10, 11, 12
30. Liu Y, Yang B, Liu Q et al. onkey: An ocr-free large multimodal model for understanding document. *arXiv:2403.04473* ; . 2, 12
31. Huang J, Yong S, Ma X et al. embodied generalist agent in 3d world. *International Conference on Machine Learning* (2024) . 2, 12
32. Hong W, Wang W, Lv Q et al. ent: A visual language model for gui agents. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 14281–14290. 2, 3, 12
33. Cherti M, Beaumont R, Wightman R et al. ducible scaling laws for contrastive language-image learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 2818–2829. 3, 4
34. Sun Q, Fang Y, Wu L et al. Eva-clip: Improved training techniques for clip at scale. *arXiv:2303.15389* ; . 4
35. Chen Z, Wang W, Tian H et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv:2404.16821* ; . 4
36. Fang Y, Wang W, Xie B et al. ring the limits of masked visual representation learning at scale. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 19358–19369. 3
37. Bavishi R, Elsen E, Hawthorne C et al. ucing our multimodal models. <https://www.addept.ai/blog/fuyu-8b> (17 October 2024, date last accessed). 3
38. Li Z, Yang B, Liu Q et al. Monkey: Image resolution and text label are important things for large multi-modal models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 26763–26773. 3
39. Liu H, Li C, Li Y et al. ed baselines with visual instruction tuning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 26296–26306. 3, 4
40. Lin Z, Liu C, Zhang R et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv:2311.07575* ; . 3
41. McKinzie B, Gan Z, Fauconnier JP et al. Methods, analysis & insights from multimodal llm pre-training. *arXiv:2403.09611* ; . 3, 4, 5
42. Elizalde B, Deshmukh S, Al Ismail M et al. Clap learning audio concepts from natural language supervision. *IEEE International Conference on Acoustics, Speech and Signal Processing* (2023) 1–5. 3
43. Girdhar R, El-Nouby A, Liu Z et al. Imagebind: One embedding space to bind them all. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 15180–15190. 3
44. Chung HW, Hou L, Longpre S et al. Scaling instruction-finetuned language models. *J Mach Learn Res* 2024; 25: 1–53. 3, 4
45. Touvron H, Lavril T, Izacard G et al. Open and efficient foundation language models. *arXiv:2302.13971* ; . 3, 4
46. Chiang WL, Li Z, Lin Z et al. : An open-source chatbot impressing gpt-4 with 90% chatgpt quality. <https://vicuna.lmsys.org> (17 October 2024, date last accessed). 3, 4
47. Touvron H, Martin L, Stone K et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288* ; . 4
48. Bai J, Bai S, Chu Y et al. technical report. *arXiv:2309.16609* ; . 3, 4, 12
49. meta. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/blog/meta-llama-3> (17 October 2024, date last accessed). 4
50. Li J, Li D, Savarese S et al. : Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, volume 202 (2023) 19730–19742. 3, 4
51. Dai W, Li J, Li D et al. blip: Towards general-purpose vision-language models with instruction tuning. *Conference on Neural Information Processing Systems* (2023) . 3, 4, 7, 8, 10
52. Liu H, Li C, Li Y et al. next: Improved reasoning, ocr, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next> (17 October 2024, date last accessed). 4, 12
53. Lu Y, Li C, Liu H et al. empirical study of scaling instruct-tuned large multimodal models. *arXiv:2309.09958* ; . 4
54. Chu X, Qiao L, Lin X et al. vlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv:2312.16886* ; . 4, 12
55. Shen S, Hou L, Zhou Y et al. e-of-experts meets instruction tuning: A winning combination for large language models. *arXiv:2305.14705* ; . 4
56. Lin B, Tang Z, Ye Y et al. ava: Mixture of experts for large vision-language models. *arXiv:2401.15947* ; . 4
57. Carion N, Massa F, Synnaeve G et al. o-end object detection with transformers. *European Conference on Computer Vision* (2020) 213–229. 4

58. Hu W, Xu Y, Li Y et al. A simple multimodal llm for better handling of text-rich visual questions. *AAAI Conference on Artificial Intelligence*, volume 38 (2024) 2256–2264. 4
59. Alayrac JB, Donahue J, Luc P et al. Fewshotgo: a visual language model for few-shot learning. *Conference on Neural Information Processing Systems*, volume 35 (2022) 23716–23736. 4
60. Wang W, Lv Q, Yu W et al. Vem: Visual expert for pre-trained language models. *arXiv:2311.03079*; . 4
61. Zhang R, Han J, Zhou A et al. Lora-adapter: Efficient fine-tuning of language models with zero-init attention. *The Twelfth International Conference on Learning Representations* (2024) . 5
62. Zeng Y, Zhang H, Zheng J et al. Why chatters in training a gpt4-style language model with multimodal inputs? *Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1 (2024) 7930–7957. 5, 8, 9
63. Yin S, Fu C, Zhao S et al. Vmbecker: Hallucination correction for multimodal large language models. *arXiv:2310.16045*; . 5, 11, 13
64. Chen L, Li J, Dong X et al. Sgpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*; . 5, 6
65. Sharma P, Ding N, Goodman S et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Annual Meeting of the Association for Computational Linguistics*, volume 1 (2018) 2556–2565. 6
66. Changpinyo S, Sharma P, Ding N et al. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021) 3558–3568. 6
67. Ordonez V, Kulkarni G and Berg T. Im2text: Describing images using 1 million captioned photographs. *Conference on Neural Information Processing Systems* 2011; 24. 6
68. Schuhmann C, Beaumont R, Vencu R et al. Im5b: An open large-scale dataset for training next-generation image-text models. *Conference on Neural Information Processing Systems*, volume 35 (2022) 25278–25294. 6
69. Schuhmann C, Köpf A, Vencu R et al. Laion coco: 600m synthetic captions from laion2b-en. <https://laion.ai/blog/laion-coco> (17 October 2024, date last accessed). 6
70. Li J, Li D, Xiong C et al. Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning* (2022) 12888–12900. 6
71. Byeon M, Park B, Kim H et al. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset> (17 October 2024, date last accessed). 6
72. Wang J, Meng L, Weng Z et al. See is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv:2311.07574*; . 6, 8, 9
73. Chen GH, Chen S, Zhang R et al. La: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv:2402.11684*; . 6, 8, 9
74. Xu J, Mei T, Yao T et al. Msr-vtt: A large video description dataset for bridging video and language. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2016) 5288–5296. 6
75. Mei X, Meng C, Liu H et al. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE ACM Trans Audio Speech Lang Process* 2024; 32: 3339–3354. 6
76. Wei J, Bosma M, Zhao VY et al. Fine-tuned language models are zero-shot learners. *International Conference on Learning Representations* (2022) . 7
77. OpenAI. Introducing chatgpt. <https://www.openai.com/research/chatgpt> (17 October 2024, date last accessed). 7
78. Ouyang L, Wu J, Jiang X et al. Aligning language models to follow instructions with human feedback. *Conference on Neural Information Processing Systems*, volume 35 (2022) 27730–27744. 7, 9
79. Sanh V, Webson A, Raffel C et al. Task prompted training enables zero-shot task generalization. *International Conference on Learning Representations* (2022) . 7
80. Zhang Y and Yang Q. Overview of multi-task learning. *Natl Sci Rev* 2018; 5: 30–43. 7
81. Gong T, Lyu C, Zhang S et al. Modal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790*; . 7, 14
82. Antol S, Agrawal A, Lu J et al. Visual question answering. *IEEE/CVF International Conference on Computer Vision* (2015) 2425–2433. 7
83. Karpathy A and Fei-Fei L. Visual-semantic alignments for generating image descriptions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2015) 3128–3137. 7
84. Xu Z, Shen Y and Huang L. Instruct: Improving multi-modal zero-shot learning via instruction tuning. *Annual Meeting of the Association for Computational Linguistics*, volume 1 (2023) 11445–11465. 7, 8, 10
85. Zhao Z, Guo L, Yue T et al. Bridge: Bridging modalities with large language model as a language catalyst. *arXiv:2305.16103*; . 8
86. Li L, Yin Y, Li S et al. M³it: Large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv:2306.04387*; . 8
87. Maaz M, Rasheed H, Khan S et al. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Annual Meeting of the Association for Computational Linguistics*, volume 1 (2023) 12585–12602. 9, 11
88. Drossos K, Liping S and Virtanen T. Clotho: An audio captioning dataset. *IEEE International Conference on Acoustics, Speech and Signal Processing* (2020) 736–740. 9
89. Wang Y, Kordi Y, Mishra S et al. Self-instruct: Aligning language model with self generated instructions. *Annual Meeting of the Association for Computational Linguistics*, volume 1 (2023) 13484–13508. 8
90. Yang R, Song L, Li Y et al. Tools: Teaching large language model to use tools via self-instruction. *Conference on Neural Information Processing Systems*, volume 36 (2023) . 8, 11, 15
91. Luo G, Zhou Y, Ren T et al. Top and quick: Efficient vision-language instruction tuning for large language models. *Conference on Neural Information Processing Systems*, volume 36 (2024) . 8

92. Wei L, Jiang Z, Huang W et al. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv:2308.12067*; . 9
93. Du Y, Guo H, Zhou K et al. Does it make sense for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *arXiv:2311.01487*; . 9
94. Ziegler DM, Stiennon N, Wu J et al. Tuning language models from human preferences. *arXiv:1909.08593*; . 9
95. Stiennon N, Ouyang L, Wu J et al. Learning to summarize with human feedback. *Conference on Neural Information Processing Systems 2020*; 33: 3008–3021. 9
96. Sun Z, Shen S, Cao S et al. Learning large multimodal models with factually augmented rlhf. *Findings of the Association for Computational Linguistics* (2023) . 10, 13
97. Rafailov R, Sharma A, Mitchell E et al. Is your preference optimization: Your language model is secretly a reward model. *Conference on Neural Information Processing Systems 2024*; 36: 10
98. Yu T, Yao Y, Zhang H et al. Towards trustworthy mlms via behavior alignment from fine-grained correctional human feedback. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 13807–13816. 10
99. Li L, Xie Z, Li M et al. Reference distillation for large visual language models. *arXiv:2312.10665*; . 10
100. Lu P, Mishra S, Xia T et al. Explain: Multimodal reasoning via thought chains for science question answering. *Conference on Neural Information Processing Systems*, volume 35 (2022) 2507–2521. 10, 14
101. Vedantam R, Lawrence Zitnick C and Parikh D. Cider: Consensus-based image description evaluation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2015) 4566–4575. 10
102. Agrawal H, Desai K, Wang Y et al. Nocaps: Novel object captioning at scale. *IEEE/CVF International Conference on Computer Vision* (2019) 8948–8957. 10
103. He X, Zhang Y, Mou L et al. Pathvqa: 30000+ questions for medical visual question answering. *arXiv:2003.10286*; . 10
104. Fu C, Chen P, Shen Y et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*; . 10, 13
105. Liu Y, Duan H, Zhang Y et al. Mmbench: Is your multimodal model an all-around player? *European Conference on Computer Vision*, volume 15064 (2024) 216–233. 10, 11
106. Ning M, Zhu B, Xie Y et al. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv:2311.16103*; . 11
107. Ye Q, Xu H, Xu G et al. Mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178*; . 11
108. Lin TY, Maire M, Belongie S et al. Microsoft coco: Common objects in context. *European Conference on Computer Vision* 740–755. 11
109. Gao P, Han J, Zhang R et al. Adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*; . 11
110. Yang Z, Li L, Lin K et al. Dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv:2309.17421*; . 11
111. Wen L, Yang X, Fu D et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving ; . 11
112. Fu C, Zhang R, Lin H et al. Challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv:2312.12436*; . 11
113. You H, Zhang H, Gan Z et al. Refer and ground anything anywhere at any granularity. *International Conference on Learning Representations* (2024) . 11
114. Lai X, Tian Z, Chen Y et al. Reasoning segmentation via large language model. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 9579–9589. 12, 15
115. Xu R, Wang X, Wang T et al. Lm: Empowering large language models to understand point clouds. *European Conference on Computer Vision* (2024) . 12
116. Sun Q, Yu Q, Cui Y et al. Integrative pretraining in multimodality. *International Conference on Learning Representations* (2024) . 12
117. Zhang D, Li S, Zhang X et al. Sogpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv:2305.11000*; . 12
118. Wang X, Zhuang B and Wu Q. Mverse: Efficiently transforming modalities with llms. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 26606–26616. 12
119. Wu S, Fei H, Qu L et al. Napt: Any-to-any multimodal lm. *International Conference on Machine Learning* (2024) . 12
120. Ho J, Jain A and Abbeel P. Using diffusion probabilistic models. *Conference on Neural Information Processing Systems 2020*; 33: 6840–6851. 12
121. Hu J, Yao Y, Wang C et al. Multilingual models pivot zero-shot multimodal learning across languages. *International Conference on Learning Representations* (2024) . 12
122. Yang Z, Liu J, Han Y et al. Agent: Multimodal agents as smartphone users. *arXiv:2312.13771*; . 12
123. Wang J, Xu H, Ye J et al. M-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv:2401.16158*; . 12
124. Ye J, Hu A, Xu H et al. M-docowl: Modularized multimodal large language model for document understanding. *arXiv:2307.02499*; . 12
125. Yu L, Poirson P, Yang S et al. Using context in referring expressions. *European Conference on Computer Vision*, volume 9906 (2016) 69–85. 12
126. Mao J, Huang J, Toshev A et al. Generation and comprehension of unambiguous object descriptions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2016) 11–20. 12
127. Zeng Y, Zhang X and Li H. Grained vision language pre-training: Aligning texts with visual concepts. *International Conference on Machine Learning*, volume 162 (2022) 25994–26009. 12
128. OpenAI. GPT-4 mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4-mini-advancing-cost-efficient-intelligence/> (17 October 2024, date last accessed). 12
129. Yao Y, Yu T, Zhang A et al. Gpt-v: A gpt-4v level llm on your phone. *arXiv preprint arXiv:2408.01800*; . 12

130. He M, Liu Y, Wu B et al. multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530* ; . 12
131. Zhai B, Yang S, Zhao X et al. switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv:2310.01779* ; . 13
132. Rohrbach A, Hendricks LA, Burns K et al. hallucination in image captioning. *Conference on Empirical Methods in Natural Language Processing* (2018) 4035–4045. 13
133. Li Y, Du Y, Zhou K et al. ating object hallucination in large vision-language models. *2023 Conference on Empirical Methods in Natural Language Processing* (2023) 292–305. 13
134. Wang J, Zhou Y, Xu G et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv:2308.15126* ; . 13
135. Liu F, Lin K, Li L et al. ing hallucination in large multi-modal models via robust instruction tuning. *International Conference on Learning Representations* (2024) . 13
136. Dong Q, Li L, Dai D et al. ey for in-context learning. *arXiv:2301.00234* ; . 13
137. Li B, Zhang Y, Chen L et al. -it: Multi-modal in-context instruction tuning. *arXiv:2306.05425* ; . 14
138. Tai Y, Fan W, Zhang Z et al. context learning for multimodal llms. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 27176–27185. 14
139. Yang X, Wu Y, Yang M et al. ring diverse in-context configurations for image captioning. *Conference on Neural Information Processing Systems*, volume 36 (2023) . 14
140. Yang X, Peng Y, Ma H et al. lm: Configuring in-context sequence to lever large vision language models. *arXiv:2312.10104* ; . 14
141. Yang Z, Gan Z, Wang J et al. empirical study of gpt-3 for few-shot knowledge-based vqa. *AAAI Conference on Artificial Intelligence*, volume 36 (2022) 3081–3089. 14
142. Lu P, Peng B, Cheng H et al. leon: Plug-and-play compositional reasoning with large language models. *Conference on Neural Information Processing Systems*, volume 36 (2023) 43447–43478. 14, 15
143. Gupta T and Kembhavi A. l programming: Compositional visual reasoning without training. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 14953–14962. 14, 15
144. Rose D, Himakunthal V, Ouyang A et al. chain of thought: Bridging logical gaps with multimodal infillings. *arXiv:2305.02317* ; . 14
145. Zhang Z, Zhang A, Li M et al. modal chain-of-thought reasoning in language models. *arXiv:2302.00923* 2023; . 14
146. Zheng G, Yang B, Tang J et al. : Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Conference on Neural Information Processing Systems*, volume 36 (2023) 5168–5191. 14
147. Ge J, Luo H, Qian S et al. of thought prompt tuning in vision language models. *arXiv:2304.07919* ; . 14
148. Parisi A, Zhao Y and Fiedel N. Tool augmented language models. *arXiv:2205.12255* ; . 14
149. Zhu X, Zhang R, He B et al. clip v2: Prompting clip and gpt for powerful 3d open-world learning. *IEEE/CVF International Conference on Computer Vision* (2023) 2639–2650. 15
150. Wang T, Zhang J, Fei J et al. n anything: Interactive image description with diverse multimodal controls. *arXiv:2305.02677* ; . 15
151. Shen Y, Song K, Tan X et al. ngpt: Solving ai tasks with chatgpt and its friends in' hugging face. *Conference on Neural Information Processing Systems*, volume 36 (2024) . 15
152. You H, Sun R, Wang Z et al. pt: Iteratively decomposing vision and language reasoning via large language models. *arXiv:2305.14985* ; . 15
153. Zeng A, Wong A, Welker S et al. tic models: Composing zero-shot multimodal reasoning with language. *International Conference on Learning Representations* (2023) . 15
154. Zhang R, Hu X, Li B et al. pt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 15211–15222. 15