Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning

Yiqi Wang,* Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao,* Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You,† Hongxia Yang

ByteDance Inc.

Abstract

Strong Artificial Intelligence (Strong AI) or Artificial General Intelligence (AGI) with abstract reasoning ability is the goal of next-generation AI. Recent advancements in Large Language Models (LLMs), along with the emerging field of Multimodal Large Language Models (MLLMs), have demonstrated impressive capabilities across a wide range of multimodal tasks and applications. Particularly, various MLLMs, each with distinct model architectures, training data, and training stages, have been evaluated across a broad range of MLLM benchmarks. These studies have, to varying degrees, revealed different aspects of the current capabilities of MLLMs. However, the reasoning abilities of MLLMs have not been systematically investigated. In this survey, we comprehensively review the existing evaluation protocols of multimodal reasoning, categorize and illustrate the frontiers of MLLMs, introduce recent trends in applications of MLLMs on reasoning-intensive tasks, and finally discuss current practices and future directions. We believe our survey establishes a solid base and sheds light on this important topic, multimodal reasoning.

Keywords Multimodal reasoning · Multimodal Large Language Model · Instruction Tuning · In-Context Learning

1 Introduction

In the past decade, with the help of increasing computational power and expanded datasets, Multimodal Large Language Models (MLLMs) have achieved remarkable progress in many domains and applications. They are considered to be the most capable family towards the ultimate goal of Strong Artificial Intelligence (Strong AI) [1] or Artificial General Intelligence (AGI) [2]. Strong AI is often thought to possess a mind, the question of whether Multimodal Large Language Models have minds, or how one might determine their existence, remains an open and complex topic. One does not need to have Sherlock Holmes' mind to reason over observations naturally in multiple modalities from the real world, such as vision, audio, text, smell, *etc.*, and subsequently take action. In the Dual-system theory [3, 4], a widespread theory of cognitive science, the second type of the human cognition systems, which enable abstract reasoning, is considered to be "*evolutionarily recent and distinctively human*". This characterization prompts an intriguing inquiry: are MLLMs capable of reasoning?

Specifically, we are interested in the reasoning tasks that require the comprehension and integration of information from various modalities, including vision, text, audio, and others. MLLMs have demonstrated effectiveness in a variety of multimodal reasoning tasks. Notable examples include Visual Question Answering (VQA) [5, 6, 7, 8], multimodal dialogue [9, 10, 11], among others. Recently, there has also been extensive research focusing on particularly improving the reasoning abilities of MLLMs, such as multimodal instruction tuning [12, 13, 14, 11] and multimodal reasoning by prompting [15, 16]. The strong capabilities of MLLMs also intrigue research on embodying them as agents in real-world environment [17, 18, 19] or enabling them using tools [20, 21]. Despite the impressive performance on

---

*Work done during an internship at ByteDance.

†Corresponding author: `quanzeng.you@bytedance.com`.

existing benchmark datasets [12, 22, 13, 23, 19, 24, 25], it is still too early to conclude that current MLLMs can truly reason over multiple input modalities.

Even in the text-only setting, Large Language Models (LLMs) are still known to lack proper reasoning abilities in some aspects, such as mathematical [26], and multi-hop reasoning [27, 28]. Moreover, both LLMs and MLLMs are also known to have hallucination issues [29], which prevent the models from reasoning properly. Motivated by the significant importance of reasoning abilities in MLLMs and the fast-paced development of related research, we believe it is necessary to comprehensively review the current status of the reasoning abilities of MLLMs and provide insightful discussion to illuminate future research. In the following sections, we will initially define the reasoning abilities of interest and illustrate existing evaluation protocols (Sec. 2); subsequently, we will summarize the current status of MLLMs (Sec. 3); this will be followed by an in-depth look at a key training phase for multimodal reasoning, namely, multimodal instruction tuning (Sec. 4); then we will discuss the reasoning-intensive applications of MLLMs, including Embodied AI (Sec. 5.1) and tool usage (Sec. 5.2); afterward, we will analyze results on multimodal reasoning benchmarks (Sec. 6). Finally, we will provide concentrated insights and discussions on the current status and future directions of MLLMs (Sec. 7).

2 Reasoning: Definition and Evaluation Protocols

2.1 What is Reasoning?

Reasoning is one of the fundamental intelligent behaviors of human beings, which requires understanding and analyzing given conditions and background knowledge to derive a new conclusion logically and rationally [30, 31, 32]. Reasoning has been extensively studied within the field of logic [33, 34, 35]. To obtain a clear understanding of reasoning, we refer to the definitions established in the field of logic [36, 37], which defines reasoning based on the concepts of premises, conclusions, and inferences. Reasoning is usually regarded as the integration of these concepts. To be more specific, premises and conclusions are true or false claims about a case. Inferences are the intermediate reasoning steps that select and interpret information from given premises, make connections, verify, and ultimately draw conclusions based on the provided and interpreted data.

Reasoning in the field of logic highly relies on mathematics [35], which is used to construct a set of basic logic rules [38]. Accordingly, only reasoning that adheres to these logical rules is considered valid. Apart from logic rules, domain knowledge is also required to perform practical reasoning tasks [39]. For instance, arithmetic reasoning necessitates mathematical knowledge, while commonsense knowledge is essential for reasoning in daily life tasks. The domain knowledge serves as additional premises besides the given inputs and is indispensable to obtaining valid conclusions in specific fields.

According to [36], reasoning can be divided into formal and informal reasoning, where conclusions of formal reasoning are guaranteed to be true as long as the premises are true, while informal reasoning does not guarantee the truth of conclusions especially when the available information is incomplete or ambiguous. In usual, informal reasoning is performed with natural language and is essential for daily-life tasks. Additionally, based on the direction of inference, reasoning can be divided into deductive, inductive, abductive, and analogical reasoning:

- **Deductive reasoning** [40] represents the most classical form of reasoning. Given a set of known knowledge (premises), it infers new knowledge step by step to obtain the conclusion. For example, given the premises that "cats are mammals" and "all mammals have four feet", deductive reasoning can infer a new conclusion that "cats have four feet". Note that deductive reasoning only concerns that the inference steps follow the logic rules and does not impose any restrictions on the truth of premises. Therefore, wrong premises may lead to wrong conclusions even if the reasoning steps are logical.

- **Inductive reasoning** concentrates on inferring general rules from specific observations [41]. For example, given the premises (observations) that "any mammal that I have seen so far has four feet", induction reasoning can infer that "all mammals have four feet". Inductive reasoning is an effective tool in scientific areas to discover new principles and laws. Note that since it is hard to collect complete observations, the conclusions of inductive reasoning may be incorrect for some unseen observations.

- **Abductive reasoning** is to infer the best explanation for the given observations [42]. It is perceived as the backward direction of deductive reasoning, where multiple reasons can lead to the results (observations) and the most likely reason should be inferred. Consider this scenario: a car is parked on the highway with its hazard lights flashing. Abductive reasoning could lead to the more plausible conclusion that the car is broken down, rather than the less likely explanation of someone playing a prank. Since the number of possible reasons is usually massive, abduction reasoning requires a lot of commonsense and domain knowledge to infer a credible reason.

- **Analogical reasoning** involves transferring knowledge from one or several instances to another, based on their similarities [43]. Two forms of analogical reasoning were studied and applied in real-life activities [44]. The first form takes as inputs one or multiple similar cases, then arrives at a hidden proposition, and finally applies the proposition to a new case. For example, consider the two cases that "iron can conduct electricity" and "copper can conduct electricity". From this, one might infer a proposition that "any metal can conduct electricity", thereby inferring that "silver, being a metal, can also conduct electricity". The second form of analogical reasoning considers the similarity of two entities to infer a property in one based on the property of the other. For instance, given the premises that "exposing plants to adequate sunlight enhances their growth and health" and "humans and plants both require certain environmental factors to thrive, like water, air, and nutrients", one could use analogical reasoning to hypothesize that "humans may also benefit from regular exposure to sunlight for their health and well-being" [45]. With analogical reasoning, the properties of a new object can be quickly inferred at a low cost. However, the premises of analogical reasoning can only support the conclusions that are likely rather than definitively true.

In this paper, we focus on the reasoning abilities of Multimodal Large Language Models. The reasoning methods employed by these models fall under the category of informal reasoning. This is primarily because they utilize natural language to articulate the steps and conclusions involved in the reasoning process and they allow a certain degree of inaccuracy in their reasoning mechanisms. This paper primarily focuses on three reasoning types: **deductive reasoning**, **abductive reasoning**, and **analogical reasoning**. These types are highlighted due to their prevalent application in real-world reasoning tasks, particularly within the scope of current MLLMs.

2.2 Language-only Reasoning Tasks

To gain a deeper insight into the reasoning abilities of Multimodal Large Language Models, it is crucial to understand the associated reasoning tasks. These tasks are widely regarded as requiring the models' reasoning capabilities for effective resolution. We can divide reasoning tasks into two categories based on the input data: language-only reasoning tasks and multimodal reasoning tasks, with the former requiring no images and the latter involving both images and text. Research on language-only tasks has a longer history, and the methodologies used for task categorization and the insights gained from these studies offer valuable guidance for the development of multimodal reasoning tasks. In this section, we will present four distinct types of language-only reasoning tasks: solving math problems, engaging in commonsense reasoning, tackling symbolic reasoning tasks, and interacting with various environments.

2.2.1 Solving Math Problems

Solving math problems usually requires one-step or multistep arithmetic reasoning. Based on the understanding of input question, the implicit arithmetic operations, and concept knowledge, a solver should infer a sequence of operation steps that can derive final answers. The range of implicit operations and conceptual knowledge can be categorized according to different school grade levels. For example, the benchmarks of GSM8K [46], SVAMP [47], ASDiv [48], and MAWPS [49] require mathematical knowledge typically acquired in primary schools. This includes fundamental operations such as addition, subtraction, multiplication, and division. The MathQA [50] benchmark and the AQuA [51] benchmark include mathematical problems sourced from standardized tests such as the GMAT (Graduate Management Admission Test) and the GRE (General Test). The MATH [52] benchmark features highly challenging mathematical problems, including areas such as permutation and combination problems, geometric series problems, solving high-order equations, among others. This benchmark requires solvers to have a lot of advanced mathematical knowledge and mathematical reasoning techniques, as well as the ability to follow a multi-step procedure, and thus remains a very challenging task.

2.2.2 Engaging in Commonsense Reasoning

Commonsense is a broadly encompassing yet somewhat loosely defined concept. While lacking exact boundaries, this generally refers to knowledge that goes beyond specialized expertise and is expected to be well-known to individuals who have completed basic education. Commonsense knowledge extends across various domains, including social common sense (*e.g.*, understanding that people would feel embarrassed if publicly blamed), physical common sense (*e.g.*, recognizing that a car is faster than a bicycle), biological common sense (*e.g.*, knowing that penguins and koalas do not naturally encounter each other), and numerous other areas. It is widely recognized that commonsense knowledge plays a significant role in everyday decision-making and real-life scenarios, making commonsense reasoning a fundamental prerequisite for language models. Recent studies have introduced several datasets aimed at commonsense reasoning, such as HellaSwag [53], Winogrande [54], Socialiqa [55], Piqa [56], CommonGen [57], Cosmos QA [58], and ART [59], which usually do not require a multi-step reasoning process.

2.2.3 Tackling Symbolic Reasoning

Symbolic reasoning can be characterized as a cognitive process carried out on abstract objects, guided by precisely defined rules like logical derivation. Besides coding and mathematical problem-solving, there exist various tasks that necessitate the application of symbolic reasoning. One such task is logistic reasoning, exemplified by datasets like PrOntoQA [60], SimpleLogic [61], FOLIO [62], and ProofWriter [63]. In these tasks, a set of facts and logical rules are given in the context and the model is required to prove a formula based on logical operations. Other tasks involve the comprehension of virtual objects. For instance, in datasets such as Penguins, Date, and Colored Objects from BIG-Bench Hard [64], there is a demand for statistical analysis and manipulation of properties related to virtual objects. An illustrative question might be, "Which penguin is named after a famous jazz musician?". While language models demonstrates the ability to understand simple symbolic manipulations, they have been identified as less adept in complex symbolic reasoning tasks [60].

2.2.4 Interacting with Various Environments

Interacting with environments require reasoning skills that involve using commonsense knowledge to understand the current situation and plan future actions. Additionally, these environments demand the ability to process feedback and adjust subsequent actions based on that feedback. This domain has seen extensive exploration in recent studies focusing on language model-driven agents. Many of these environments primarily employ textual modalities as their basis. Examples include Textworld [65], Alfworld [66], and WebShop [67]. These virtual scenarios rely on textual descriptions for state representation, feedback, and interaction. In these environments, agents are tasked with completing various objectives. For instance, they might be required to place fresh lettuce on a dining table within a living room setting or to choose a 3-ounce bottle of citrus-scented deodorant from a simulated online shopping platform. Beyond purely textual environments, some environments incorporate multiple modalities to challenge language model reasoning. For instance, BabyAI [68] and MineDojo [69] present game-like scenarios. Within these environments, models are expected to engage in tasks such as crafting an iron sword in the Minecraft game. Furthermore, a subset of studies has delved into the realm of reasoning within real-world contexts. For example, Ahn *et al.* [70] tackled 101 robotic tasks within an actual kitchen, pushing the boundaries of language model reasoning beyond simulated scenarios.

2.3 Multimodal Reasoning Benchmarks

With the ever-evolving multi-modal large language models' reasoning capability, benchmarks play a pivotal role in assessing their performance, understanding their capabilities, and identifying areas that demand improvement. Based on the aforementioned definition and categorization of reasoning, an ideal multimodal reasoning benchmark is supposed to 1) truly require multimodal information; 2) follow the categorization of reasoning; and 3) have detailed annotation of reasoning steps to examine whether the reasoning process is correct.

2.3.1 Benchmark Datasets

Traditional datasets, such as COCO caption [71], Nocaps [72], and Flickr30K [73], are utilized to evaluate the MLLMs' comprehension of image content and captioning. Datasets focused on visual question answering, such as VQAv2 [5], OK-VQA [6], ScienceQA [7], and GQA [8], feature question-answer pairings. These datasets serve as an initial platform for evaluating a model's reasoning capabilities. Recently, more evaluation benchmarks have been introduced to holistically compare the various capabilities of MLLMs. We summarize existing evaluation benchmarks in Table 1, with an emphasis on their reasoning-related evaluation.

Surprisingly, except for the recently introduced Infi-MM-Eval [74], most existing reasoning-related multimodal benchmarks are not reasoning-focused and do not follow the aforementioned criteria of an ideal benchmark. For example, another manually labeled benchmark MM-Vet [75] does not follow the reasoning categorization to properly evaluate different reasoning types, nor contain reasoning steps to examine the reasoning process. Another example is MMMU [76], a dataset collected from textbooks. It does not specify the type of reasoning required and only includes partial annotations of the reasoning steps. For simplicity and to ensure our empirical findings are more focused on reasoning and robustness, our analysis will primarily concentrate on these three datasets in Sec. 6.

2.3.2 Evaluation Metrics

With these rich evaluation datasets shown in Table 1, it is important to evaluate using proper metrics. Therefore, it is necessary to categorize and review the evaluation metrics. The complexity of evaluating MLLMs arises from the multimodal nature of these tasks. The metrics must capture the synergistic relationship between language and vision, as well as other possibly involved modalities. This relationship demands assessments that not only estimate the accuracy of predictions but also the depth, nuance, and relevancy of multimodal associations.

Table 1: Evaluation Benchmark Summarization. Benchmarks are ordered by releasing date.

Dataset	Description	Stats	Reasoning Part	Tags
MMMU[76]	Designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning	11.5K questions from college exams, quizzes, and textbooks		Manual Collect, Multiple-Choice Accuracy
InfiMM-Eval[74]	Designed for open-ended complex reasoning QA with intermediate steps	279 high quality samples		Manual Collect, GPT eval, Full reasoning step
HallusionBench[77]	benchmark for evaluating both visual and language hallucination	346 images with 1129 questions		Manual Collect, Public Source, GPT eval
MathVista[78]	Consolidated Mathematical reasoning benchmark with visual contexts, covers 7 types of reasoning, e.g. algebraic, arithmetic, geometry, logical, numeric common sense and scientific	6,141 samples		Manual Collect, Public Source, Multiple-Choice Accuracy
LLMDoc[79]	OCR-free document instruction understanding evaluation set, including table, chart, document, natural image, and webpage	100 samples		Manual Collect, Public Source, Human eval
PCA-Eval[80]	Evaluating the decision-making ability of embodied agents from three perspectives: perception, cognition, and action. Images are from various embodies environments	300 multi-choice questions	each question has a "reason" field for evaluating reasoning capability	Manual Collect, Multiple-Choice Accuracy
SparklesEval[11]	GPT-assisted benchmark for assessing a model's conversational competence across multiple images and dialogue turns	150 dialogs, total 550 images		GPT Generate, GPT eval
M3Exam[81]	Multilingual, multimodal, multilevel, multiple choice questions from exam papers	12137 questions, 23% contain images		Manual Collect, Multiple-Choice Accuracy
TouchStone[9]	Comprehensive visual dialogue dataset, consisting of open-world images and questions, covering 5 categories of abilities and 27 subtasks	908 questions	comprehension questions at 29.6%, 3.6% for multi-image analysis	Manual Collect, GPT eval
MM-Vet[75]	QAs for 6 VL capabilities with 187 online images and 13 public dataset images	218 questions	11.9% of math and 34.4% of spatial awareness	Manual Collect, GPT eval
MMBench[82]	Vision-language QAs in the format of multiple choice over 20 abilities	2,974 questions	1114 Reasoning Questions	Manual Collect, Multiple-Choice Accuracy
MME[83]	Yes and No choice questions on 14 subtasks. Each image has 2 questions	1097 images	130 images	Manual Collect, Multiple-Choice Accuracy
LAMM-Benchmark[84]	9 image tasks and 3 point cloud tasks	62439 images and 12788 point cloud samples, 186k instruction-response pairs	ScanQA, SQAimage and AI2D	GPT Generate, Traditional Metrics, GPT eval
CCEval[85]	benchmark for evaluating detailed caption hallucination with object coverage	100 VisualGenome images		Public Source, GPT eval
VisIT-Bench[86]	Instruction following evaluation with human-authored instruction-conditioned caption	592 queries with 1159 images		Manual Collect, GPT eval, Human eval
SEED-Bench[87]	Multiple choice questions, spans 12 evaluation dimensions in images and videos. Questions generated by GPT, answered by human	19K questions	331 visual reasoning, 97 instance interaction and 657 spatial relations	Manual Collect, GPT Generate, Multiple-Choice Accuracy
I4[14]	Evaluating instruction following ability on interleaved VL instructions, covers 19 scenarios with 29 tasks	18k instructions with 62k images		Public Source, Multiple-Choice Accuracy, Caption Score
LVLM-eHub[88]	Combination of many public datasets and an online arena platform	47 visual benchmarks	DocVQA, OKVQA, ScienceQA, SNLI-VE etc.	Public Source, Human eval, Caption Score, Multiple-Choice Accuracy
OwlEval[25]	50 collected images with questions, some has multi-turn conversations	82 questions		Manual Collect, Human eval
Open-VQA[89]	Open-ended questions including images and videos	450 questions	31 Reasoning Questions	Manual Collect, Human eval
ScienceQA[90]	Multimodal multiple-choice questions with annotations and explanations covers natural science, social science, and language science, collected from online education platform	21208 questions, 10332 contain images		Manual Collect, Multiple-Choice Accuracy
Winoground[91]	Measuring visio-linguistic compositional reasoning	1600 image-text pairs, comprise 400 examples		Manual Collect, Multiple-Choice Accuracy
RAVEN[92]	Multi-choices Raven's Progressive Matrices (RPM) questions	1,120,000 images and 70,000 RPM problems		Rule Generate, Multiple-Choice Accuracy

The existing evaluation metrics can be categorized from different aspects. Although we mainly focus on three datasets for results analysis, we categorize all the related datasets for a comprehensive summarization. From the perspective of evaluating the different capabilities of models, evaluation can be categorized into 3 types:

- For *vision capability*, traditional perception metrics are relevant, including tasks such as identifying, classifying, and locating objects within an image, as well as interpreting those objects in the context of the overall scene. The common metrics used are image classification accuracy (ImageNet [93]), object detection mAP (COCO [94]), object segmentation mIoU (LVIS [95]), and others.

- For *linguistic capability*, commonsense understanding and coherence needed to be considered as metrics. It goes beyond merely generating grammatically correct sentences; it's about producing responses that are contextually relevant, logical, and semantically coherent, ensuring alignment with the given visual input. The common metrics used are BLEU [96], CIDEr [97], ROUGE [98], among others.

- For *reasoning capability*, it covers spatial reasoning, knowledge reasoning (including areas like common sense, mathematics, text, code, and more), as well as hypothesis-based reasoning. Just as humans deduce information based on textual and visual cues, MLLMs should exhibit the ability to reason, draw inferences, and predict outcomes based on the multi-modal data they process. The common metrics are QA-Accuracy [76], Elo score [99], GPT-4 [100] evaluation, and others.

From the types of evaluation questions, existing multimodal benchmarks can be divided into two distinct categories:

- ***Closed-Set Evaluation Benchmarks*** are circumscribed by a predetermined set of categories or outcomes. When MLLMs are assessed under this paradigm, they're essentially provided a limited set of potential responses or classifications. Such evaluations gauge the model's aptitude to correctly categorize or respond within the given confines, making them particularly suited for tasks where boundaries are well-defined and the space of responses is finite. Certain visual evaluation tasks are of a closed-set nature, such as image classification and attribute classification. Conversely, some benchmarks with open-ended questions force to reformat answers into multiple-choice format, examples of which includes MMBench [82], MME [83] and ScienceQA [90]. However, as studied in a recent work [101], authors found that popular MLLMs are vulnerable to adversarial permutation in answer sets for multiple-choice prompting.

- ***Open-Set Evaluation Benchmarks*** are more unbounded and exploratory. They allow models to generate responses without confining them to a predefined set. This evaluates the model's capability to venture into the unknown, generating or deducing outcomes that might not be part of the training set. In the vast and dynamic domain of multimodal interactions, where real-world scenarios often elude strict categorizations, open-set evaluations become crucial in assessing a model's generalization capabilities and adaptiveness to novel inputs. Two common approaches for evaluating open-set answers are human scoring such as LVLM-eHub [88] and Lynx [89], and automatic language model scoring, like TouchStone [9] and VisIT-Bench [86].

With the three datasets subject to our analysis, diverse reasoning evaluation metrics are employed, including GPT-4 evaluation (Infi-MM-Eval, MM-Vet) and QA-Accuracy (MMMU). Additionally, these datasets feature different evaluation question setups: Open-Set (Infi-MM-Eval, MM-Vet) and Closed-Set (MMMU). Therefore, a representative subset of benchmarks are covered in our analysis.

2.4 Improving Reasoning Abilities for LLMs

Since reasoning abilities are essential for solving complex tasks, a wealth of research has focused on improving reasoning abilities in LLMs. Using supervised data to pretrain or fine-tune LLMs is a straightforward method to enhance their reasoning capabilities. For instance, during the pretraining stage, Lewkowycz *et al.* [102] incorporate a variety of mathematical corpora into their training set to improve quantitative reasoning abilities. In the fine-tuning stage, Rajani *et al.* [103] utilize a small instruction-tuning dataset to fine-tune the pretrained GPT model. Beyond supervised training, in-context learning [104, 105] and prompt engineering [106] have also shown great potential for improving reasoning abilities. such as the Chain-of-Thought approach [104] and its variants [107, 108]. Another line of research [109, 110, 111] has proposed interacting LLMs with external tools, such as a Python code interpreter [111], to enhance the accuracy of each reasoning step. Within the domain of MLLMs, these methods, namely instruction-tuning, prompt engineering, and tool usage have also been explored for multimodal reasoning, which we will later elaborate.

3 Multimodal Large Language Models

3.1 Overview

Multimodal Large Language Models typically comprise a visual encoder and a language model, linked by a connector. Both the visual encoder and the language model are commonly initialized from pretrained models. The distinctions among MLLMs can be analyzed based on the following key aspects:

Training Data: Multimodal training data can be segmented into three parts. **(1)** Pretraining data is utilized for aligning other modalities, such as the visual modality, with the Large Language Model. **(2)** (Optional) High-quality multi-task supervised learning data comprises question-answer pairs used to inject knowledge into the model. **(3)** The instruction tuning data enhances Multimodal Large Language Models by improving their abilities to follow instructions and engage effectively in user interactions.

V-L Interaction Module: As shown in Figure 1, some models incorporate visual tokens as if they were a foreign language, directly injecting them into the input layer. For instance, LLaVA [13] directly inputs the visual signal into the input layer of the large language model. In contrast, other models, such as Flamingo [12], employ cross-attention layers to facilitate interactions between visual and language features within transformer blocks.
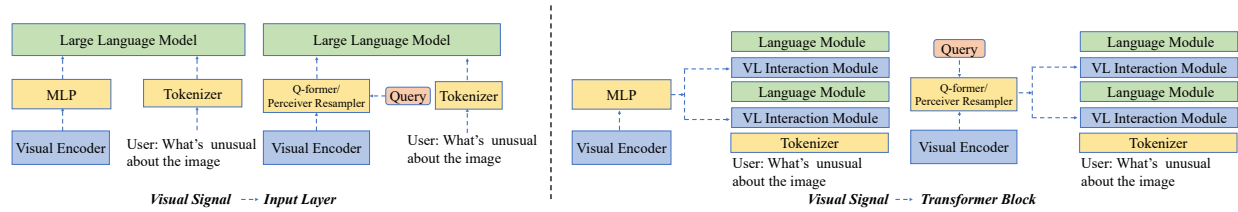

Figure 1: Various architectures of Multimodal Large Language Models.

Connector Architecture: The specifics of the connector design also play a pivotal role in determining the capabilities of MLLMs. A key differentiating factor among these models is their choice of visual-language connectors. Models such as BLIP-2 [22], Flamingo [12], and QWen-VL [112] utilize query-based connectors like Q-former/perceiver resampler. Conversely, LLaVA [13] and MiniGPT4-v2 [113] employ a multilayer perceptron (MLP) as the connector.

3.2 Large Language Models

Recent advancements in MLLMs represent an evolution of LLMs, which makes reviewing of LLM development both relevant and beneficial. GPT-3 [114] pioneered this domain, showcasing a powerful emergent capability in zero-shot generation tasks. This was evident in its in-context learning and chain-of-thought processes, achieved using 175B parameters and pretraining on extensive web-text data. Subsequently, ChatGPT [115] further augmented these capabilities by incorporating instruction tuning and reinforcement learning from human feedback. GPT-4 is an advanced iteration of ChatGPT; however, its training specifics remain undisclosed due to the source code not being made public. Other noteworthy LLMs, such as PaLM [116], BLOOM [117], Chinchilla [118], LLaMA [119], OPT [120], GLM [121], Alpaca [122] and Vicuna [123], have also been introduced, propelling the domain of LLMs forward.

3.3 Multimodal Large Language Models

We give a brief timeline in Figure 2, the efficacy of Multimodal Large Language Models is enhanced by incorporating vision encoders and connectors. This integration allows for training on diverse multimodal datasets. Table 2 highlights key distinctions among various MLLMs. The Flamingo architecture [12] pioneers the usage of query-based cross-attention mechanism, termed the "perceiver resampler". This innovative approach constructs a robust vision-language interactive module, showcasing remarkable abilities in in-context learning. BLIP-2 [22], integrates the Q-Former. This query-based sampler, initialized from pretrained BERT, aims to bridge the gap between vision and language modalities.

These methods demonstrate their proficiency in generating text from images, a capability achieved through various stages of pretraining alignment. Later, InstructBLIP [124], enhances the pretrained BLIP-2 by training on instruction tuning datasets sourced from diverse public datasets. This augmentation results in improvements across the majority of zero-shot performance metrics, surpassing the performance of BLIP-2. LLaVA [13] further enhances the instruction-following ability of MLLMs by training on visual instruction tuning data. Its subsequent iteration, *LLaVA-1.5* [125], expands upon this by integrating VQA datasets into the instruction tuning data, leading to significant performance

improvements across a range of benchmarks. Notably, both InstructBLIP and LLaVA-1.5 follow a two-stage training process, which includes pre-training and subsequent instruction tuning.

In contrast, QWen-VL [112] introduces a more complex three-stage training process. This process includes pretraining, multi-task training, and instruction tuning. A particular emphasis is placed on the use of high-quality supervised data during the multi-task training stage. Additionally, QWen-VL attempts to fine-tune all modules of the model, including the vision encoder, connector, and the large language model.

Furthermore, MiniGPT-v2 [113] also employs a three-stage framework. This framework emphasizes task-specific instruction templates for diverse tasks and involves fine-tuning both the projection layer and the Large Language Model, showcasing considerable improvements over MiniGPT4. LLaMA-Adapter [23] incorporates a lightweight adaptation method for fine-tuning. This method excels in instruction-following and multimodal reasoning tasks, particularly image-based question answering. Additionally, mPLUG-owl2 [25] leverages modality collaboration. This approach enhances performance in both text-only and multimodal tasks. Meanwhile, Otter [126] refines the OpenFlamingo model, which focuses on improving adherence to instructions and effectively utilizing in-context samples. Notably, the CogVLM [127] demonstrates superior performance by integrating an extra visual expert in the language model, achieving state-of-the-art performance on several benchmarks.
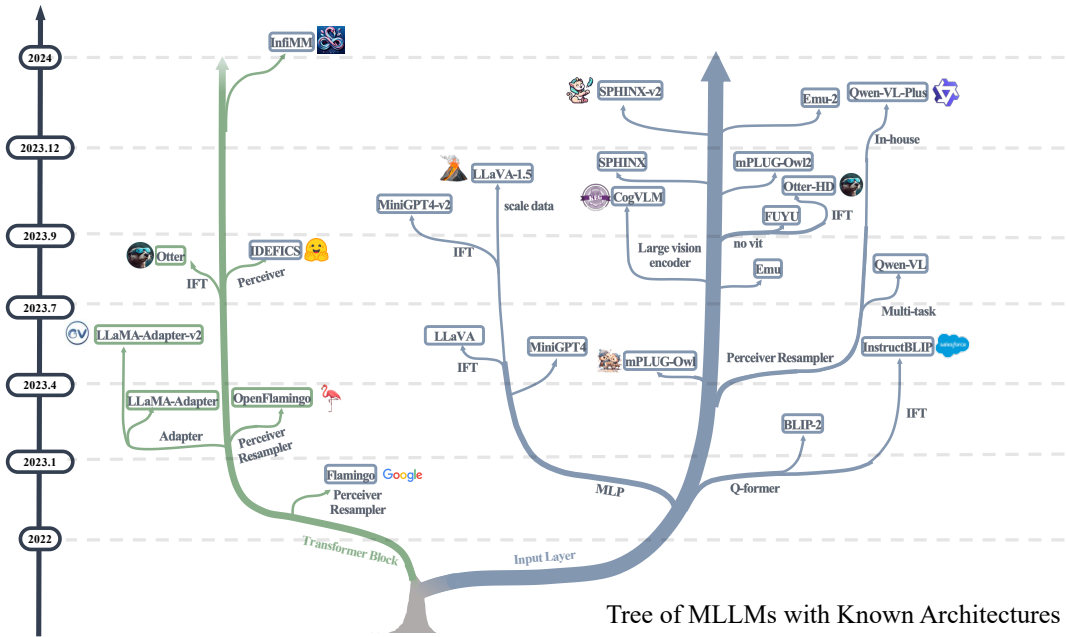

Figure 2: A brief timeline outlining recent developments in MLLMs.

4 Multimodal Reasoning through Instruction Tuning

Pre-trained LLMs are capable of complex reasoning in a few-shot manner when prompted with task-related exemplars or rationales [128, 104, 106]. One of the methods to achieve such ability of in-context learning (ICL) is through instruction tuning. With appropriate tuning, the model is expected to enhance its ICL capabilities, *e.g.* when prompted with examples of input-output pairs. However, there are several main challenges when the prompting modalities go beyond language. In this section, we first review current works on instruction tuning that facilitate ICL, and then the primary challenges and recent solutions in multimodal prompting.

4.1 Definition of Instruction Tuning

In recent years, instruction tuning has attracted significant attention due to the remarkable achievements of Large Language Models such as GPT-3[128], LLAMA [119], Claude [129], PaLM [116]. These models have demonstrated exceptional capabilities in various human-level understanding and reasoning tasks, including mathematics [52, 46], coding [130], and most notably in generalizing to new tasks with zero or few-shot examples through instruction tuning.
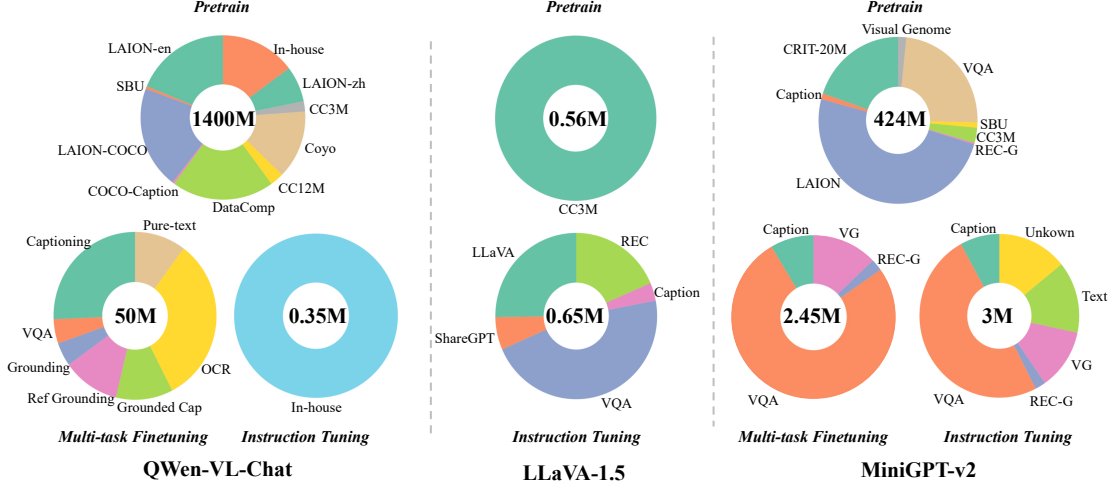
Figure 3: The datasets used in different stages of various Multimodal Large Language Models are depicted. At the center of each circle, the total number of samples (in millions) for each stage is displayed.

Pretrained LLMs often struggle with generalizing to new tasks and aligning with users' intents, leading to untruthful, unhelpful, and unsafe responses. To mitigate these limitations, instruction tuning [131, 115] has been proposed.

Instruction tuning typically begins with a carefully collected dataset comprising thousands of ⟨*instruction*, *response*⟩ pairs. In this process, LLMs are fine-tuned using this dataset to better align human intent with model behavior. It is worth noting that the tuning of model parameters is supervised by the loss function $\mathcal{L}_{it}$, which is computed based solely on the ***response*** tokens. The loss function is defined as:

$$\mathcal{L}_{it} = - \sum_{x_i \in \mathbf{response}} \log p(x_i | x_{0:i-1}, \mathbf{instruction}). \tag{1}$$

The pursuit of high-quality and diverse instruction-tuning datasets has been a key goal. A well-crafted ***instruction*** usually includes clearly described question that represents human intent, along with any necessary background question context. In addition, the ***response*** must be accurate and detailed. In the context of multi-modal instruction tuning, each example in the dataset consists of ⟨***X-modality***, ***instruction***, ***response***⟩, where ***X-modality*** typically represents images, videos, or audio. Table 2 shows current literature and their training dataset, with some of these involving instruction tuning stage.

4.2 Instruction Tuning for In-Context Learning

In-context learning on Large Language Models has gained significant attention as a crucial indicator of models' reasoning capabilities. Models excelling in ICL should inherently possess strong analogical reasoning ability. LLMs with robust ICL capability can adapt to new tasks based on just a few examples, without the need for updating their weights. Recently, Multimodal Large Language Models built upon LLMs have also demonstrated such ICL capability. Frozen [15] is the first MLLM to showcase ICL ability using a frozen LLM. Flamingo [12] demonstrated strong ICL capabilities with a more powerful LLM and web-scale images interwoven with text for pretraining. OpenFlamingo [157] and IDEFICS [158] are open-source reproductions of Flamingo. While they may not match Flamingo in every performance metric, they have demonstrated similar In-Context Learning (ICL) capabilities.

All ICL-capable MLLMs mentioned above are pretrained models. During the instruction tuning stage, most MLLMs primarily focus on image-text pairs, which can gradually lead to a diminishing of ICL capabilities. This is attributed to the challenges in constructing ICL-focused instruction tuning datasets. Recently, Otter [126] has developed the image-text-interleaved instruction tuning dataset MIMIC-IT [159], which was then fine-tuned on OpenFlamingo. Another notable instruction fine-tuning dataset is MIC [160], which includes in-context examples and is composed of 16 public datasets. In MIC, samples and in-context examples are constructed by randomly sampling from pre-defined templates. While these advancements have equipped MLLMs with ICL abilities, challenges continue to exist in the area of multimodal prompting.

Table 2: Recent multimodal large language models,

Methods	Visual Encoder	Pre-training Stage	Multi-task learning Stage	Instruction Tuning
OpenFlamingos	CLIP-ViT-L/14 [132]	LAION-2B [133], Multimodal C4 [134], Synthetic data	-	-
LLaMA-Adapter-v2	CLIP Vit-L/14	GPT-4-LLM [135], COCO	-	-
BLIP2	CLIP ViT-L/14	COCO [94], CC3M [136], CC12M [137], LAION400M [138], Visual Genome [139],	-	-
InstructBLIP	ViT-G/14	COCO, Visual Genome CC3M [136], CC12M [137], LAION400M [138]	-	26 publicly available vision-language datasets, transformed into the instruction tuning format
LLaVA-1.5	CLIP ViT-L/14	LLaVA [13]	-	LLaVA665k (VQAv2 [140], GQA [8], OK-VQA [6], AOK-VQA [141], OCRVQA [142], TextCaps [143], LLaVA150k [13], ShareGPT [144])
mPLUG-owl2	CLIP Vit-L/14	LAION-400M, COYO [145], COCO, Laion-en, DataComp [146]	-	VQAv2, OKVQA, OCR-VQA, GQA, A-OKVQA, RefCOCO [147], Visual Genome, LLaVA150K, ShareGPT, SlimOrca [148]
MiniGPT-v2	EVA-G	GRIT-20M [24], LAION, CC3M, SBU [149], COCO caption, Text Captions [143], RefCOCO [147], RefCOCO+, RefCOCOg [150], Visual Genome, GQA, VQAv2, OCR-VQA, OK-VQA, AOK-VQA	GQA, VQAv2, OCR-VQA, OK-VQA, AOK-VQA, RefCOCO, RefCOCO+, RefCOCOg, Visual Genome, COCO caption, Text Captions	COCO Caption, Text Captions RefCOCO, RefCOCO+, RefCOCOg, Visual Genome, GQA, VQAv2, OCR-VQA, OK-VQA, AOK-VQA, LLaVA150k, Flickr30k [73], Multi-task conversation,Unnatural Instructions
CogVLM	EVA2-CLIP-E	LAION-2B, COYO	REC	LLaVA150k, LLaVAR [151], LRV-Instruction [152], in-house data
QWen-VL-Chat	CLIP ViT-bigG	LAION-en, LAION COCO, Data-Comp, Coyo, CC12M, CC3M, SBU, COCO Caption, LAION-zh, In-house Data	LAION, DataComp, Coyo, CC12M, CC3M, SBU, COCO, GQA, VGQA, VQAv2, DVQA [153], OCR-VQA, DocVQA [154], TextVQA, ChartQA [155], AI2D [156], GRIT, Visual Genome, RefCOCO, Ref-COCO+, RefCOCOg, SynthDoG, Common Crawl, In-house Data	In-house data

4.3 Multimodal Prompting by Representation Learning

Multimodal prompting, while opening up avenues to leverage rich information from non-linguistic modalities, often leads to catastrophic forgetting [161] due to disparities in information densities and a lack of high-quality paired data. A key challenge in this domain is to derive LLM-compatible representations from non-linguistic inputs without altering the base LLM. This challenge has spurred research in multimodal representation learning, aiming to develop a unified prompt embedding space for various modalities. As an attempt to accommodate a secondary prompting modality and to harness the knowledge from the pre-trained LLM, Tsimpoukelli *et al.* [15] train a vision encoder alongside a frozen LLM to transform vision prompts into an embedding sequence compatible with language embeddings. Additionally, Koh *et al.* [162] proposed a method for learning linear projections to convert embeddings between vision and language. This approach enables their contextual-retrieval-based model FROMAGe to uniformly handle multimodal prompts.

4.4 Multimodal Prompting by Exemplar Generation

Prompting with in-context examples is a paradigm that unlocks the analogical reasoning ability of the pretrained LLMs for downstream tasks [163]. Although this approach is widely used in single-modality contexts [128, 164, 165, 166], there are relatively few studies on prompting with multimodal examples. A significant challenge in multimodal in-context learning is that multimodal LLMs often require contextual information to tackle new tasks, but acquiring in-context examples for every new task is not always feasible. Addressing this issue, Guo *et al.* [167] presented Img2LLM, a system capable of automatically generating LLM-agnostic exemplar prompts for visual question answering based on question images. This is achieved by first extracting answer candidates from generated captions and then formulating questions that correspond to these answers, thus creating a set of question-answer pairs suitable for new tasks.

4.5 Multimodal Prompting by Model Interaction

Combining the distinct capabilities of foundation models across different modalities for general multimodal tasks remains challenging, often requiring computationally expensive joint pretraining or fine-tuning. Zeng *et al.* [16] introduced the concept of multimodal prompting. This approach facilitates the exchange of information between
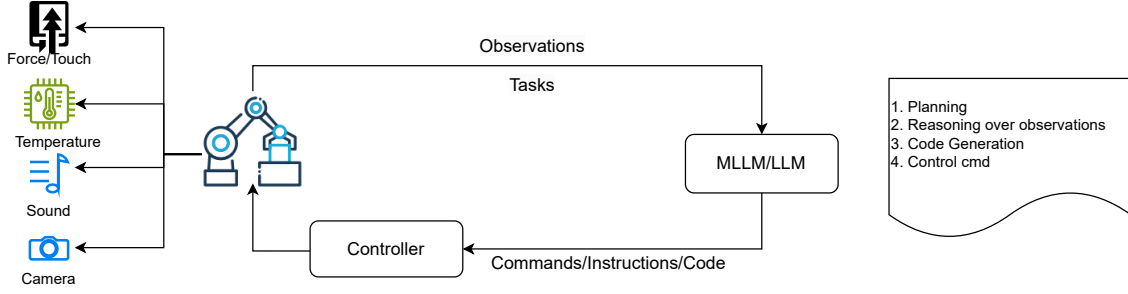
Figure 4: Illustration of the interaction between an Embodied Agent and MLLM/LLM models. Here, pretrained models generate commands and planning outcomes for the robot. In turn, the robot provides environmental observations to the MLLM models, aiding in refining their reasoning for future tasks.

different modalities through language prompts. They demonstrate that by thoughtfully crafting interactive processes and prompt templates for specific multimodal tasks, these foundation models can perform new tasks in a zero-shot manner, eliminating the need for additional training. Wang *et al.* [168] similarly uses text as the interface to realize multimodal in-context learning without any training.

5 Applications

In this section, we mainly cover Embodied AI and Multimodal Agent to discuss the applications of MLLM models.

5.1 Embodied AI

Embodied AI has become increasingly popular with the advancement of powerful AI models. The Embodied AI workshop at the conference on Computer Vision and Pattern Recognition (CVPR) offers this definition:

> *Minds live in bodies, and bodies move through a changing world. The goal of embodied artificial intelligence is to create agents, such as robots, which learn to creatively solve challenging tasks requiring interaction with the environment.*

> *https://embodied-ai.org/*

This field requires the integration of multiple technologies and knowledge areas, including Computer Vision, Speech, Natural Language Processing, and Control Systems. These are essential for enabling an agent to see, talk, listen, and interact with its environment. More importantly, it also demands that the agent possess reasoning abilities, allowing it to plan actions based on environmental cues and adapt its behavior in response to feedback from its surroundings.

Robotics stands as the foremost representation of embodiment AI agents. The recent advancements in MLLMs shed light on robotic manipulation by leveraging the commonsense knowledge from the visual world and the reasoning ability of LLMs. Figure 4 illustrates the interaction loop between an Embodied Agent and an MLLM. Specifically, this commonsense knowledge assists the Embodied agents in tasks like planning, complex reasoning, code generation, and the creation of control commands. These generated instructions, in various formats, are then provided to the control unit for the generation and execution of actions. This section provides a summary of the latest applications of MLLMs in embodied AI, with a focus on their reasoning capabilities.

5.1.1 Tasks and Challenges of Embodied AI Reasoning

Research tasks in Embodied AI are broadly classified into three categories: visual exploration, visual navigation, and embodied question answering [169, 170]. These tasks require the embodied agent to perceive its environment, plan and execute control strategies, and interact with humans or the environment. A primary challenge in this field is adapting agents to out-of-domain environments, which demands robust commonsense knowledge and reasoning abilities from the agent. This aligns with the recent advancements in pretrained large models, including both LLMs and MLLMs. The subsequent subsections will delve into how MLLMs can be applied to enhance the capabilities of Embodied Agents.

5.1.2 Direct Reasoning with MLLM Knowledge

Early methods of employing MLLMs primarily centered on the direct use of pre-trained LLM and MLLM models. These models typically incorporate commonsense knowledge acquired from extensive datasets. Observations gathered by the Embodied Agent are provided to these models, leveraging their pretrained reasoning abilities for task planning and problem-solving. The principal strategy involves decomposing the embodied tasks into smaller, more executable steps, allowing the pre-trained model to efficiently tackle each intermediate stage. Initially, these studies focused less on additional training of the model and more on the task decomposition and the efficient coordination of subtasks.

Reasoning with Separate Components. CLIPort [171] represents an early endeavor to harness the semantic knowledge of pretrained models trained on large-scale datasets. In particular, it utilizes CLIP [132] to assist broad semantic understanding. Featuring a two-stream architecture, their end-to-end framework is capable of solving various tasks without the need for explicit concept learning for each task.

Later, Socratic Models (SMs) are proposed in [16], where each pretrained model is assumed to contain commonsense knowledge from distinct domains. By combining these models, they can collectively address tasks across various domains. Specifically, different models can be composed by using language prompts as the bridge. The proposed Socratic Models employ a multimodal prompting strategy to combine different VLM, LM models.

In a related study, LM-Nav [172] leverages pretrained models for robotic navigation. The main idea is to employ language as the bridge to break down the instructions. The process involves employing GPT-3 for identifying textual landmarks for instructions, using CLIP for landmark grounding, and optimizing a probabilistic objective by inferring a joint likelihood over landmarks and nodes. Again, this method requires no further model fine-tuning.

More recently, Dasgupta *et al.* [173] explored the Embodied AI agent concept within a **Planner-Actor-Reporter** paradigm. The "Planner" interprets human instructions and directly translates them into sequences of simple steps to be executed by the "Actor", which can be implemented by a pretrained LLM. The "Reporter" observes the environment, gathers information, and reports it to the "Planner". The information will be appended to the previous instructions to form the new prompt. Their experiments highlight the advantages of this paradigm. They also suggest that the "Reporter" can be trained within a Reinforcement Learning (RL) framework.

Reasoning with External World Model. Currently, LLM has a limit of window context. To execute a task that involves the states or facts from the distant past, this work proposes to use a world state to represent the knowledge [174]. In particular, the proposed approach attempts to use memory or a world model to record the key states of the current environment. The mechanism includes a world state reader and world state writer, which utilizes the states as prompts for LLM to solve a complex task and update the world state memory according to new observations. Experimental results highlight the effectiveness of this model, surpassing existing methods in certain aspects. However, there are areas for improvement. The method still requires task-specific prompt definition, lacks visual feedback, and the world state may be updated inaccurately.

Reasoning with Feedback. Inner Monologue [175] attempts to imitate human thought processes in *natural languages*. Inspired by human cognition, this work proposes a method for solving robotic tasks by using languages as a bridge. Robots are equipped with "short-horizon skills", enabling them to execute basic tasks or commands described in simple language. Additionally, a Large Language Model is employed as a planner. This LLM aims to fulfill high-level user commands by breaking them down into a sequence of skills that the robot can perform.

Furthermore, the system can also incorporate environmental "textual" feedback into its prompts, a process described as an "inner monologue". This feedback includes success detection, object and scene description, Visual Question Answering (VQA), and human feedback. The authors highlight the emergent capabilities of their system. Most of the reasoning and re-planning are facilitated by the LLM. In particular, the authors attribute the success of this reasoning to the incorporation of environment feedback into the model's decision-making process.

Reasoning with Policy Generation. The research presented in [176] further studies the application of language model programming for robotic control. Moving beyond the direct use of the pretrained LLM's language capabilities, this study also involves prompting an LLM with coding ability to generate policy code.

This approach further expands the scope of reasoning tasks to include arithmetic and novel instructions. Specifically, this study uses OpenAI Codex code-davinci-002. It uses prompts, supplemented with examples, to guide the LLM in generating additional functions that adhere to user instructions.

Subsequently, the generated code is executed using the robot's control primitives. As discussed in the paper, this system effectively utilizes LLMs and code to enhance its reasoning capacity. However, it is constrained by the limitations of

the available API and control primitives. Additionally, the system lacks a mechanism to assess the feasibility of a given instruction.

5.1.3 Improved Reasoning with Robotic data

Collecting robot trajectories with crowd-sourced annotations is challenging, often resulting in datasets of limited scale. The authors from [177] propose data driven instruction augmentation for language-conditioned control (**DIAL**). Their approach consists of several key steps: 1) Fine-tuning a Vision-Language Model (VLM) on a small, manually annotated dataset of robot manipulation trajectories; 2) Gathering a larger, unannotated dataset; 3) Using the fine-tuned VLM to annotate this newly collected dataset; 4) Retraining the VLM on both the original and the newly labeled dataset.

In the re-labeling step, the authors utilize GPT-3 to expand the range of human instruction candidates. They explore various strategies for selecting appropriate instructions for given robot trajectories. Overall, DIAL demonstrates more capabilities in solving novel tasks. It can also be used to distill Internet-scale vision language models into embodied agents [178].

Another approach involves integrating discrete text tokens directly into the training set as natural language tokens [17], further enriching the model's training data.

5.1.4 Embodied Foundation Model

PaLM-E [19]: the Largest Reported Embodied Multimodal Large Language Model. The primary goal is to transform inputs from various modalities into formats compatible with LLMs, thereby enabling embodied agents to interact with environments across different modalities. The paper evaluates various methods for converting these modalities into LLM inputs. In particular, the object scene representations [179] demonstrate its effectiveness even with limited training data in robotic applications. The largest PaLM-E-562B demonstrates emergent abilities, including zero-shot multimodal chain-of-thought reasoning. This is largely due to its ability to transfer knowledge from general visual-language understanding to tasks related to embodied AI.

However, the reasoning or planning skills of PaLM-E are integrated with separate low-level policies that generate robot control commands for execution. In this work, the authors employ RT-1 [180] for these low-level policies. RT-1 processes visual observations and language instructions, converting them into robot control commands.

RT2: Vision-Language-Action (VLA) Model. Subsequently, the following approach, RT-2 [17], directly utilizes a pretrained VLM model to generate robot control commands. To achieve this, they apply the action discretization pipeline. This pipeline transforms each dimension of the seven-dimensional continuous action space into 256 uniformly distributed discrete actions.

Following this transformation, the RT-2 model undergoes fine-tuning with both robotics data and original web data. This approach shows better generalization than models fine-tuned solely with robotic-only data.

For the base VLM, the authors employ two of the largest VLM models available: PaLI-X [18] and PaLM-E [179]. Their results prove that after fine-tuning the VLA models with robot trajectory data, RT-2 demonstrates several emergent abilities. These include enhanced reasoning capabilities with novel objects and improved interpretation of human instructions. More interestingly, after only a few hundred gradient steps of fine-tuning, RT-2 acquires chain-of-thought reasoning abilities. The data augmentation process introduces an additional "Plan" step, wherein the model is required to plan its actions before executing them. This seemingly simple addition significantly enhances the model's ability to comprehend and respond to more complex commands.

5.2 MLLMs for Tool Usage

5.2.1 Tool Usage in LLM

Recent studies have explored the potential of using external tools to augment LLMs' capabilities for complex tasks [111]. Research [181, 182, 183] has demonstrated that supplementing LLMs with tools, such as calculators, search engines, translation systems, calendars, or even API calls to other models can help solve tasks beyond their inherent capabilities.

The integration of tools into LLMs generally involves two primary phases: tool selection or creation, and processing the results obtained from these tools [184, 185, 186]. Tool selection, a pivotal phase within tool learning, can involve allowing the LLM itself to make decisions based on input queries [187, 188, 189, 111, 19]. Alternatively, it might include the use of predefined tools specifically designed for particular scenarios[182, 190, 191, 192], such as retrieval systems.

When the model is required to autonomously determine which tools to employ, it necessitates the model to understand the characteristics of these tools and make accurate selections via their reasoning abilities. The majority of approaches tend to decompose a complex task into a series of subtasks interacting with tools in sequential or tree structures, such as Chain of Thoughts (CoT) [104, 106], Tree of Thoughts (ToT) [193] and Program of Thoughts (PoT) [110].

A simple way to teach language models to use tools is by instruction [111, 191, 183]. The description of tools is able to teach language models to use tools [191] in a zero-shot fashion. Language models can also learn to use tools from few-shot examples. This in-context tool learning can range from retrieval tasks to enhancing the model's ability to generalize to new challenges. For example, Art [183] demonstrates this by selecting examples of multistep reasoning and tool usage from a task library. This approach significantly improves upon traditional few-shot prompting and automated Chain of Thought (CoT) methods, particularly in unseen tasks within the BigBench and MMLU benchmarks.

Some studies have shown that LLMs can learn to utilize tools by fine-tuning on the tool-specific corpora [181, 182, 194]. TALM [181] and Toolformer [182] both adopt a self-supervised method, starting with a few tool demonstrations and then enabling the language model to autonomously generate additional data involving tool usage. TALM is designed to use APIs for search and calculation, applying these functions to enhance language models' capabilities in question-answering tasks. Toolformer, on the other hand, utilizes a broader range of tools, including a calculator, a QA system, a search engine, a translation system, and a calendar, leading to improved zero-shot performance across various downstream tasks.

Gorilla [194] introduces APIBench, a comprehensive dataset comprising APIs from HuggingFace, TorchHub, and TensorHub. The study demonstrates that a fine-tuned LLaMA-based model, trained with this dataset, significantly surpasses GPT-4 in generating API calls, showcasing its enhanced performance.

Recent research [188, 110] has expanded the capabilities of LLMs beyond just utilizing existing tools; it now includes enabling LLMs to create tools through programming. This method uses programming languages to tackle mathematics tasks, thereby enhancing the models' capacity for accurate mathematical computation and improving logical task planning.

5.2.2 MLLMs as Tool

Table 3: Comparison of Works Using MLLM as Tools. For Modalities Integrating Works, we list the tool domain instead of enumerating all individual tools.

Application	Work	Reasoning method	Tools
Compositional Visual Reasoning	VISPROG [195]	Program	OWL-ViT, DSFD, MaskFormer, CLIP, ViLT, Stable Diffusion
Multimodal Dialogue	Visual ChatGPT [20]	CoT	BLIP, Stable Diffusion, Pix2Pix, ControlNet
	MM-ReAct [196]	ReAct	image captioning, tagging, celebrity recognition, OCR, Web
Modalities Integrating	Chameleon [190]	Program	Modalities: image, web, knowledge, math, table
	TaskMatrix.AI [197]	CoT	Modalities: image, office, cloud, robot

MLLMs can function as tools to empower language models in handling various modalities and addressing challenging visual reasoning tasks. We briefly review these works from different application perspectives, as illustrated in Table 3.

Compositional Visual Reasoning. While current MLLM models demonstrate proficiency in visual tasks such as image classification and captioning, they encounter difficulties with more complex visual operations, like visual editing, due to inherent requirements for reasoning and planning. For instance, replacing a person's face with a cat's face in an image involves face detection, person recognition, cat face generation, and merging the new face into the original image. Critical reasoning is required to break down the tasks and develop a detailed execution plan. MLLMs can play a pivotal role in carrying out these decomposed steps.

To solve compositional visual reasoning tasks, VISPROG [195] utilizes MLLMs as a toolset within a modular and interpretable neuro-symbolic system. This system, designed for comprehensive visual understanding and manipulation, leverages GPT-3 to generate programs through in-context learning. These programs coordinate different modules, such as OWL-ViT [198] for open vocabulary localization, DSFD [199] for face detection, MaskFormer [200] for segmentation, CLIP [132] for image selection and classification, ViLT [201] for visual question answering, and Stable Diffusion [202] for object replacement. Additional tools such as Python Imaging Library[3] (Pillow) and OpenCV[4] are also used for image manipulation. VISPROG can handle various complex visual reasoning tasks, including

---

[3] https://pypi.org/project/pillow/
[4] https://opencv.org/

compositional VQA, zero-shot reasoning on image pairs, factual knowledge object tagging, and image editing using natural language.

Multimodal Dialogue Reasoning.   Following dialog instructions can become more complex than single-step instructions due to their context-dependent nature. Multi-turn dialogue understanding requires reasoning based on dialog history. MLLMs can serve as tools for LLMs to facilitate dialogs that incorporate visual inputs and outputs, drawing on the reasoning capabilities of LLMs to interpret dialog instructions.

Visual ChatGPT [20], incorporating various MLLM, enables users to interact with ChatGPT using both text and images. This system harnesses MLLMs such as BLIP [22], Stable Diffusion [202], and other visual foundation models like Pix2Pix [202], ControlNet [203]. It is capable of performing tasks like image captioning, image question answering, and image editing during conversations. Chain-of-Thought (CoT) [104] reasoning method is used for case that requires collaboration between multiple visual models.

Additionally, MM-ReAct [196] utilizes the ReAct [111] planning method alongside a suite of image and video tools for a stronger capacity of visual tasks. These tools cover a wide range of functions, including image captioning, image tagging, dense captioning, celebrity recognition, receipt scanning, OCR, and Bing search. This enables MM-ReAct to handle difficult tasks such as Visual Math and Text Reasoning, Visual-Conditioned Jokes/Memes, and Visual Planning.

Integrating Modalities Beyond the Visual Domain.   MLLMs are usually customized for specific modalities, such as images, which poses a challenge when trying to integrate their usage across varied domains like mathematics, the web, tables, and more. For instance, solving a math problem depicted in an image requires the visual model to first interpret the image and then invoke the math model for a solution. Addressing this challenge entails fostering reasoning and planning among various modality models. By conceptualizing all the modality models as tools and leveraging the reasoning capabilities of LLM, recent efforts aim to integrate additional modalities beyond the visual domain, such as web content, tables, office documents, and robotic data. This approach contributes to the development of a more powerful multimodal model.

Chameleon [190] exemplifies this approach. It can utilize a variety of tools, such as existing vision models, web search engines, Python functions, and heuristic-based modules. Capable of handling tasks across image, web, knowledge, math, and table domains, Chameleon synthesizes programs by combining these tools to coordinate their sequence through an LLM-driven planner for multistep tasks.

TaskMatrix.AI [197] provides a comprehensive set of APIs catering to a wide range of tasks. These include image editing, image-based question answering, image captioning, text-to-image conversion, and transformations like image-to-sketch/depth/hed/line. Furthermore, TaskMatrix.AI extends its utility to conversational robotics and IoT device control, offering APIs for tasks like robotics pick, move, and put, car air conditioner control, TV operation, and music playback. TaskMatrix.AI utilizes the CoT [104] reasoning method to device tool usage plans.

6   Multimodal Reasoning Benchmark Result Analysis

Table 4: Evaluation results for various MLLM on MMVet and MMMU (validation set). Open-source models best performances are indicated with underlines.

MLLMs	MMVet	MMMU
MiniGPT-4 [113]	24.4	26.8
OpenFlamingo-v2 [157]	24.8	28.7
LLaMA-Adapter V2 [204]	31.4	29.8
CogVLM-Chat [127]	**52.8**	32.1
Otter [126]	24.6	32.2
mPLUG-Owl2 [205]	36.3	32.7
BLIP-2 [22]	22.4	35.4
InstructBLIP [124]	26.2	35.7
QWen-VL-Chat [112]	-	35.9
LLaVA-1.5 [13]	36.3	36.4
InfiMM-LLaMA-13B [206]	39.2	**39.1**
GPT-4V [100]	**67.7**	**55.7**

Table 5: Evaluation results for various MLLMs on InfiMM-Eval. The top performances achieved by open-source models are indicated with underlines for easy reference.

MLLMs	Reasoning Category			Reasoning Complexity		Overall
	Deductive	Abductive	Analogical	Moderate	High	
OpenFlamingo-v2 [157]	8.88	5.3	1.11	9.47	4.72	6.82
MiniGPT-v2 [113]	11.02	13.28	5.69	14.45	7.27	10.43
Fuyu-8B [210]	16.42	21.49	7.78	23.06	9.91	15.7
BLIP-2 [22]	22.76	18.96	7.5	24.05	14.18	19.31
InternLM-XComposer-VL [211]	26.77	35.97	18.61	39.13	17.18	26.84
InstructBLIP [212]	27.56	37.76	20.56	40.64	18.09	28.02
LLaMA-Adapter V2 [204]	28.7	46.12	22.08	41.33	21.91	30.46
Otter [126]	22.49	33.64	13.33	35.79	12.31	22.69
mPLUG-Owl2 [205]	23.43	20.6	7.64	28.79	13.18	20.05
IDEFICS-9B-instruct [213]	22.99	34.63	20.56	34.45	16.73	24.53
Emu [214]	28.9	36.57	18.19	36.18	22.0	28.24
LLaVA-1.5 [13]	30.94	47.91	24.31	47.4	21.0	32.62
CogVLM-Chat [127]	36.75	47.88	28.75	55.67	22.5	37.16
Qwen-VL-Chat [112]	37.55	44.39	__30.42__	46.61	__30.09__	37.39
SPHINX-v2 [215]	__42.17__	49.85	20.69	54.85	27.31	39.48
InfiMM-LLaMA-13B [206]	39.9	__53.18__	__31.94__	__55.85__	28.7	__40.7__
GPT-4V [100]	**74.86**	**77.88**	**69.86**	**93.98**	**58.98**	**74.44**

In this section, we briefly summarize the performance of existing Multimodal Large Language Models on the evaluation benchmarks outlined in Sec. 2.3. Our focus is primarily on accessing the reasoning capabilities of these models.

Table 4 presents a summary of the results from both MM-Vet and MMMU datasets. MM-Vet [75] is an open-ended QA benchmark designed for evaluating the comprehensive capabilities of MLLMs. According to their leaderboard, GPT-4V has achieved the best score among all MLLMs. The GPT-4 driven MM-ReAct [21] also performs exceptionally well, particularly in reasoning tasks like mathematics and spatial reasoning. This indicates that MLLMs' reasoning capabilities are heavily influenced by their underlying LLM. Among the open-source models, CogVLM-Chat has achieved the highest score. Compared with other public models, the response of IFT dataset used in CogVLM-Chat are mostly long format and go through further manual correction. Such data aligns better with open-ended evaluation, thus achieved highest score on MM-Vet.

MMMU [76] is a benchmark designed to evaluate MLLMs on college-level subject questions. The leaderboard reveals intriguing results: proprietary MLLMs, like Gemini Ultra/Pro [207], GPT-4V [208], and Qwen-VL-PLUS [209] achieved much better scores than most open-source MLLMs. Notably, the scores of these open-source MLLMs are even lower than those of LLMs aided by OCR tools and image captions generated by LLaVA. MLLMs that reserve LLMs' capability, like Blip2/Instruct-Blip and InfiMM-LLaMA-13B achieved good performance, suggesting that MMMU requires model to retain knowledge obtained during LLM training and most other open-source MLLMs may experience catastrophic forgetting of language capability after visual IFT.

InfiMM-Eval[74] is an open-ended question-answering benchmark designed to evaluate MLLMs' visual reasoning capability. In contrast to MMMU, which heavily relies on the knowledge capacity of the underlining LLM, InfiMM-Eval focuses more on the fundamental unified logic capabilities of MLLMs. The evaluation results on InfiMM-Eval are shown in Table 5. Similar to other benchmarks, GPT-4V has shown a significant lead over other MLLMs. Among the open-source models, SPHINXv2 [215], Qwen-VL-Chat [112], and CogVLM-Chat [127] rank as the top three. While Qwen-VL-Chat and CogVLM-Chat have achieved similar overall scores, their performance varies across different levels of reasoning complexity. Since CogVLM-Chat employs similar underline LLMs with other models, it is hypothesized that Qwen-VL-Chat's superior performance in highly complex questions may stem from the Qwen language model, whereas CogVLM-Chat's strengths in moderately complex questions are likely due to its enhanced detailed perception capability.

In reviewing the top-performing open-source models on reasoning-focused benchmarks such as SPHINX-v2, QWen-VL-Chat, CogVLM-Chat and InfiMM-LLaMA-13B, we summarize the following common recipes to facilitate future research:

Table 6: Evaluation results for MLLMs before and after instruction tuning on InfiMM-Eval.

MLLMs	Reasoning Category			Reasoning Complexity		Overall
	Deductive	Abductive	Analogical	Moderate	High	
Qwen-VL-7B [112]	24.06	24.39	9.03	32.33	12.69	21.32
Qwen-VL-7B-Chat [112]	32.25	48.18	24.03	46.90	22.78	33.44

- Unfreezing the language model at a certain training stage is instrumental in enhancing multimodal reasoning abilities. For example, SPHINX-v2 and QWen-VL-Chat unfreeze their language models during the pretraining and instruction tuning stages, respectively. CogVLM, on the other hand, trains visual experts to modify the language model's response to visual input and also unfreeze the language model during fine-tuning. While tuning the language model is expected to improve multimodal performance, careful control of training parameters such as the number of training steps and the learning rate is crucial to prevent catastrophic forgetting.

- Improving visual representations is key to better multimodal performance. This can be achieved by increasing the resolution of images for finer details, as done in QWen-VL-Chat and CogVLM, or by employing stronger or multiple visual encoders, like in SPHINX-v2. Another effective method is fine-tuning the visual encoder during the MLLM training process, as demonstrated by QWen-VL-Chat.

- The multi-task supervised learning stage significantly contributes to the superior performance of top models like QWen-VL-Chat, CogVLM, and InfiMM-LLaMA-13B. Each of these models employs a three-stage training process, demonstrating the value of this approach. Notably, InfiMM-LLaMA-13B, while using the same instruction dataset as LLaVA-1.5, showcases improved reasoning abilities across various benchmarks. The multi-task supervised learning stage can be achieved by utilizing datasets of higher quality than the pretraining stage, yet containing less conversational data compared to the instruction tuning stage. Additionally, CogVLM achieves impressive performance on the mm-vet, utilizing unknown in-house data during the instruction tuning stage.

Besides these general strategies, we also observed dataset-specific trends and variances:

- As shown in Table 6, instruction tuning significantly enhances performance on InfiMM-Eval, arguably the most representative dataset for evaluating multimodal reasoning. The importance of instruction tuning for MLLMs is also verified by the fact that MLLMs usually take instruction tuning as their final training stage.

- There are significant differences in model rankings between InfiMM-Eval and MMMU. For example, BLIP-2, which performs relatively poorly on InfiMM-Eval, excels on MMMU, surpassing models such as CogVLM. Due to a range of factors, including the types of evaluation datasets, language model architectures, and training methodologies, it is challenging to pinpoint the exact cause of these discrepancies. However, it is important to highlight the potential for performance gaps across various evaluation datasets.

Note that our analysis is more of a case-study approach, focusing on the findings from top-performed models, especially in the context of the reasoning-focused InfiMM-Eval dataset. Therefore, these insights may not be universally applicable across various training data source, language models, and their respective scales.

7 Conclusions and Future Directions

The concept of reasoning ability is pivotal in the quest for Strong Artificial Intelligence (Strong AI) or Artificial General Intelligence (AGI), a goal that has been pursued across various scientific disciplines for several decades. With recent advancements in both models and evaluation benchmarks, there is a growing discourse on the reasoning abilities exhibited in current LLMs and MLLMs. In this work, we present the different types of reasoning and discuss the models, data, and evaluation methods used to measure and understand the reasoning abilities demonstrated in existing studies. Our survey aims to provide a better understanding of our current standing in this research direction and hopes to inspire further exploration into the reasoning abilities of future work.

In addition, we wish to outline some potential research directions to improve the reasoning capabilities of current MLLMs. The following discussions include our summaries regarding pretraining, alignment and evaluation benchmarks:

MLLM Architectures. Current MLLM architectures still have the following fundamental limitations in achieving human-level multimodal reasoning ability. The first limitation comes from the image resolution. Current Multimodal

Large Language Models typically operate on images with a fixed, preset resolution. However, in real-world scenarios, images can come in a variety of resolutions. The hallucination problem of MLLM also possibly stems from the MLLM architecture design. With additional modalities as input, it is more complicated to diagnose the source of the hallucinations in MLLMs. It becomes challenging to determine whether these hallucinations arise from an incomplete perception of visual signals or from biases and spurious correlations ingrained in language models. Therefore, it is important to design MLLM architectures that can inherently minimize or avoid hallucinations caused by inaccurate perception.

Efficiency and Scalability of MLLM Training Recipes. As shown in Figure 3, different MLLMs utilize vastly varying amounts of data in each training stage. However, performance comparisons do not show proportional differences (Table 5 and Table 4). Understanding the relationship between data volume and model scale at each training stage is crucial. According to recent studies [216, 125, 217], the pretraining stage for aligning vision-text representation may not require extensive computational resources.

Long-context Support. Existing MLLMs often serve in short-context situations and lack the ability to handle diverse real-world, long-context scenarios. These scenarios range from comprehending lengthy documents and papers to more complex tasks like understanding videos or movies. While most MLLMs employ attention mechanisms, the challenge with long-context tasks is the need to minimize computational complexity and maintain focus as the prompt length increases. A significant gap exists in research on multimodal long-context tasks, largely attributable to the absence of appropriate datasets and evaluation benchmarks that are equipped to tackle these specific challenges.

Instruction Fine-tuning (IFT) Data. An enhanced approach to instruction fine-tuning data is necessary, given that most of the existing public MLLM instruction datasets are insufficient in developing reasoning abilities for multimodal tasks. Currently, there is a lack of ablation studies exploring the type of data that could boost multimodal reasoning capabilities. Therefore, the creation of future instruction datasets demands careful consideration to more effectively improve reasoning capabilities in multimodal contexts.

Reinforcement Learning for Multimodal Reasoning. Inspired by the success of InstructGPT [115], leveraging reinforcement learning algorithms [218] to align model predictions and human preference has been extensively studied [129, 219]. However, reinforcement learning with human feedback (RLHF) has been just applied to MLLM [220] and there are many exciting open research problems. For example, given the complexity of multiple modalities, it is important to collect preference data effectively and scalably. With sufficient data, it is also unclear whether existing learning algorithms can handle the multimodal tasks or not. Another example is how to develop multimodal agents capable of actively engaging and utilizing tools for complex, reasoning-intensive tasks.

Evaluation Benchmarks. Current evaluation benchmarks focus on single-round conversations, revealing a gap for multi-round, multi-image conversational benchmarks. Such benchmarks require models to respond to a variety of instructions within a single conversation, which could significantly advance research in enhancing models' generalization capabilities and reducing over-fitting to specific instructions.

References

[1] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

[2] Ben Goertzel and Cassio Pennachin. *Artificial general intelligence*, volume 2. Springer, 2007.

[3] Jonathan St BT Evans. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459, 2003.

[4] Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[7] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

[8] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[9] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models, 2023.

[10] Shengzhi Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs, 2023.

[11] Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Collier Nigel, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*, 2023.

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

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

[14] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions. *arXiv preprint arXiv:2308.04152*, 2023.

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

[16] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[17] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[18] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023.

[19] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

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

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

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[23] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention, 2023.

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

[25] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

[26] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.

[27] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[28] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[29] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[30] Fei Yu, Hongbo Zhang, and Benyou Wang. Nature language reasoning, a survey. *arXiv preprint arXiv:2303.14725*, 2023.

[31] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[32] Douglas N Walton. What is reasoning? what is an argument? *The journal of Philosophy*, 87(8):399–419, 1990.

[33] Craig DeLancey. *A concise introduction to logic*. Open SUNY Textbooks, 2017.

[34] Maurice A Finocchiaro. Informal logic and the theory of reasoning. *Informal Logic*, 6(2), 1984.

[35] Edward H Madden. Reasoning and the logic of things: The cambridge conferences lectures of 1898. *The Philosophical Review*, 103(2):380–382, 1994.

[36] Hugo Bronkhorst, Gerrit Roorda, Cor Suhre, and Martin Goedhart. Logical reasoning in formal and everyday reasoning tasks. *International Journal of Science and Mathematics Education*, 18:1673–1694, 2020.

[37] Bradley H Dowden. Logical reasoning. 2018.

[38] Vladimir V Rybakov. *Admissibility of logical inference rules*. Elsevier, 1997.

[39] Jian-Bo Yang, Jun Liu, Jin Wang, How-Sing Sii, and Hong-Wei Wang. Belief rule-base inference methodology using the evidential reasoning approach-rimer. *IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans*, 36(2):266–285, 2006.

[40] Philip N Johnson-Laird. Deductive reasoning. *Annual review of psychology*, 50(1):109–135, 1999.

[41] Ming Li and Paul MB Vitanyi. Inductive reasoning and kolmogorov complexity. *Journal of Computer and System Sciences*, 44(2):343–384, 1992.

[42] Igor Douven. Abduction. 2011.

[43] Usha Goswami. Analogical reasoning: What develops? a review of research and theory. *Child development*, 62(1):1–22, 1991.

[44] Paul Bartha. Analogy and analogical reasoning. 2013.

[45] Merrilee H Salmon. *Introduction to logic and critical thinking*. Cengage Learning, 2012.

[46] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[47] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021.

[48] Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online, July 2020. Association for Computational Linguistics.

[49] Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California, June 2016. Association for Computational Linguistics.

[50] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*, 2019.

[51] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*, 2017.

[52] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[53] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019.

[54] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740, 2020.

[55] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, 2019.

[56] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[57] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, 2020.

[58] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, 2019.

[59] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2019.

[60] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2022.

[61] Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. *arXiv preprint arXiv:2205.11502*, 2022.

[62] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022.

[63] Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, 2021.

[64] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[65] Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pages 41–75. Springer, 2019.

[66] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

[67] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

[68] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: A platform to study the sample efficiency of grounded language learning. In *International Conference on Learning Representations*, 2018.

[69] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

[70] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[71] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[72] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[73] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

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

[75] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023.

[76] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2023.

[77] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models, 2023.

[78] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint Link*, 2023.

[79] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding, 2023.

[80] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *ArXiv*, 2023.

[81] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. 2023.

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

[83] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[84] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, 2023.

[85] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. Halle-switch: Controlling object hallucination in large vision language models, 2023.

[86] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use, 2023.

[87] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.

[88] Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, and Ping Luo. Tiny lvlm-ehub: Early multimodal experiments with bard, 2023.

[89] Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs?, 2023.

[90] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[91] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022.

[92] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[93] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[94] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[95] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019.

[96] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[97] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.

[98] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[99] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[100] OpenAI. Gpt-4 technical report, 2023.

[101] Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations, 2023.

[102] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

[103] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019.

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

[105] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

[106] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[107] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022.

[108] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

[109] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR, 2023.

[110] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, 2022.

[111] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023.

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

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

[114] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

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

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

[117] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model, 2023.

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

[119] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

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

[121] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland, May 2022. Association for Computational Linguistics.

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

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

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

[125] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

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

[127] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.

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

[129] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[130] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[131] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[132] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[134] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.

[135] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[136] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[137] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[138] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

[139] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[140] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[141] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[142] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[143] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.

[144] sharegpt.

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

[146] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

[147] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[148] Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023.

[149] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[150] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[151] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[152] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1, 2023.

[153] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

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

[155] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[156] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688, 2021.

[157] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[158] Hugo Laurencon, Daniel van Strien, Stas Bekman, Leo Tronchon, Lucile Saulnier, Thomas Wang, Siddharth Karamcheti, Amanpreet Singh, Giada Pistilli, Yacine Jernite, and Victor Sanh. Introducing idefics: An open reproduction of state-of-the-art visual language model, August 2023.

[159] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning, 2023.

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

[161] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[162] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. *ICML*, 2023.

[163] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[164] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.

[165] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[166] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1423–1436, Toronto, Canada, July 2023. Association for Computational Linguistics.

[167] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023.

[168] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497, 2022.

[169] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[170] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[171] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[172] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pages 492–504. PMLR, 2023.

[173] Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. Collaborating with language models for embodied reasoning. *arXiv preprint arXiv:2302.00763*, 2023.

[174] Takuma Yoneda, Jiading Fang, Peng Li, Huanyu Zhang, Tianchong Jiang, Shengjie Lin, Ben Picker, David Yunis, Hongyuan Mei, and Matthew R Walter. Statler: State-maintaining language models for embodied reasoning. *arXiv preprint arXiv:2306.17840*, 2023.

[175] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782. PMLR, 2023.

[176] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[177] Ted Xiao, Harris Chan, Pierre Sermanet, Ayzaan Wahid, Anthony Brohan, Karol Hausman, Sergey Levine, and Jonathan Tompson. Robotic skill acquisition via instruction augmentation with vision-language models. *arXiv preprint arXiv:2211.11736*, 2022.

[178] Theodore Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. Distilling internet-scale vision-language models into embodied agents. *arXiv preprint arXiv:2301.12507*, 2023.

[179] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in Neural Information Processing Systems*, 35:9512–9524, 2022.

[180] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[181] Aaron Parisi, Yao Zhao, and Noah Fiedel. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*, 2022.

[182] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023.

[183] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*, 2023.

[184] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models, 2023.

[185] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: a survey, 2023.

[186] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities, 2023.

[187] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023.

[188] Cheng Qian, Chi Han, Yi R. Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. Creator: Disentangling abstract and concrete reasonings of large language models through tool creation, 2023.

[189] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022.

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

[191] Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Tool documentation enables zero-shot tool-usage with large language models, 2023.

[192] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022.

[193] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

[194] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

[195] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023.

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

[197] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*, 2023.

[198] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*.

[199] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019.

[200] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

[201] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[202] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[203] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[204] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023.

[205] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023.

[206] InfiMM Team. Infimm:, 2024.

[207] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[208] OpenAI. Gpt-4v(ision) technical work and authors., 2023.

[209] Qwen-VL Team, 2023.

[210] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023.

[211] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023.

[212] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[213] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023.

[214] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[215] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

[216] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7005–7015, 2021.

[217] Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14846–14855, 2023.

[218] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[219] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[220] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.