

MMAT-1M: A Large Reasoning Dataset for Multimodal Agent Tuning

Tianhong Gao* Yannian Fu*† Weiqun Wu Haixiao Yue Shanshan Liu Gang Zhang
Baidu Inc.

Abstract

Large Language Models (LLMs), enhanced through agent tuning, have demonstrated remarkable capabilities in Chain-of-Thought (CoT) and tool utilization, significantly surpassing the performance of standalone models. However, the multimodal domain still lacks a large-scale, high-quality agent tuning dataset to unlock the full potential of multimodal large language models. To bridge this gap, we introduce MMAT-1M, the first million-scale multimodal agent tuning dataset designed to support CoT, reflection, and dynamic tool usage. Our dataset is constructed through a novel four-stage data engine: 1) We first curate publicly available multimodal datasets containing question-answer pairs; 2) Then, leveraging GPT-4o, we generate rationales for the original question-answer pairs and dynamically integrate API calls and Retrieval Augmented Generation (RAG) information through a multi-turn paradigm; 3) Furthermore, we refine the rationales through reflection to ensure logical consistency and accuracy, creating a multi-turn dialogue dataset with both Rationale and Reflection (RR); 4) Finally, to enhance efficiency, we optionally compress multi-turn dialogues into a One-turn Rationale and Reflection (ORR) format. By fine-tuning open-source multimodal models on the MMAT-1M, we observe significant performance gains. For instance, the InternVL2.5-8B-RR model achieves an average improvement of 2.7% across eight public benchmarks and 8.8% on the RAG benchmark Dyn-VQA, demonstrating the dataset’s effectiveness in enhancing multimodal reasoning and tool-based capabilities. The dataset is publicly available at <https://github.com/VIS-MPU-Agent/MMAT-1M>.

1. Introduction

In recent years, Multimodal Large Language Models (MLLMs) exemplified by GPT-4o [31], Gemini [35], the QwenVL series [1, 2, 39], the InternVL series [8–10], and the LLaVA series [21, 22] have made remarkable strides. To further enhance the reasoning and problem-solving capabilities

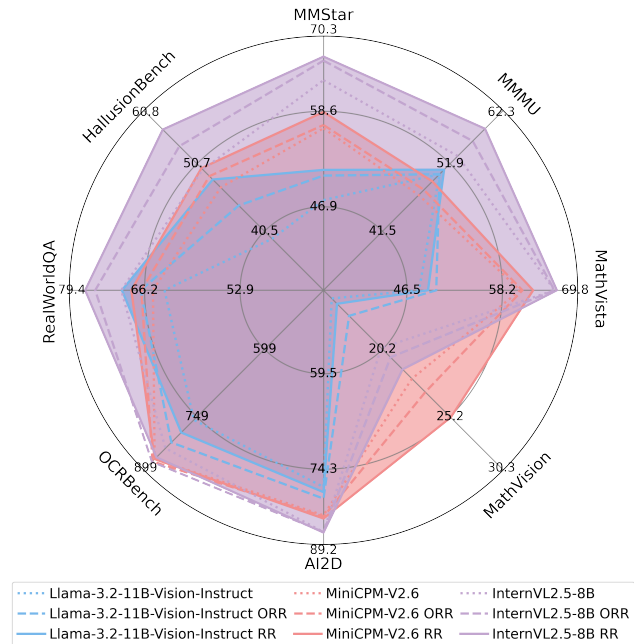


Figure 1. Performance comparison of multimodal large language models fine-tuned on MMAT-1M dataset using One-turn Rationale and Reflection (ORR) and Rationale and Reflection (RR) across eight benchmarks. Both strategies significantly boost performance, demonstrating the effectiveness of structured reasoning and MMAT-1M.

of these models, integrating Chain-of-Thought (CoT) reasoning and external tools has proven to be an effective approach, commonly referred to as “Agents”. Agents operate through two primary methods: instruction-driven [13, 33, 40, 44, 45] and tuning-driven [3, 7, 36, 47, 49]. The former involves designing prompts to enable LLMs to plan, reason, and utilize tools, which demands strong prompt comprehension. The latter employs specialized datasets to fine-tune models, empowering even smaller models to achieve agent capabilities comparable to proprietary large models. Consequently, agent tuning has emerged as a prominent and promising research direction.

In terms of existing research, several representative works have emerged in the field of multimodal agent tun-

*Equal contribution.

†Corresponding author.

ing. For instance, LLaVA-Plus [23] converts LLaVA-158K dataset into a tool-use instruction format with 117K samples through both user-oriented and skill-oriented dialogues, and T3-Agent [14] constructs the MM-Traj dataset that contains 20K multimodal tasks with tool-usage trajectories. However, existing datasets commonly suffer from three critical shortcomings: (1) They exhibit a relatively homogeneous distribution, limiting improvements to diverse benchmarks; (2) They lack mechanisms for reflecting on errors induced by visual tools, resulting in weak model robustness against interference; (3) They are deficient in flexible reasoning and tool-usage mechanisms, reducing their feasibility for real-world applications. Consequently, building a large-scale tuning dataset that effectively addresses these challenges—diversity, robustness, and flexibility—has emerged as a critical breakthrough for advancing the field.

To overcome these bottlenecks, we propose Multi-Modal Agent Tuning—One Million (MMAT-1M), which, to the best of our knowledge, is the first million-scale multimodal agent tuning dataset including diverse fundamental visual tasks. Building on publicly available multimodal datasets, we design a four-stage data synthesis framework. First, we compile publicly accessible multimodal datasets that encompass question-answer pairs. To ensure consistency in input and output formats across diverse multimodal datasets, we adapt the prompts for both inputs and outputs. Then we generate iterative trajectories using CoT reasoning and dynamic API calls, incorporating functionalities such as Image Caption, Optical Character Recognition (OCR), Open-Vocabulary Object Detection (OVD), Face Detection, and RAG. Next, we evaluate these trajectories for logical inconsistencies and refine those requiring modification through a reflection process. To enhance practical flexibility, we optionally consolidate iterative trajectories into a one-turn format and prepend tool-usage results to the input. Experimental results demonstrate that models fine-tuned with the MMAT-1M dataset exhibit significant performance advantages. As Figure 1 shows, after training on our two formats of datasets, all three mainstream open-source models achieve better performance compared to the baseline. Taking the InternVL2.5-8B-RR model as an example, it achieves an average improvement of 2.7% across eight publicly available multimodal benchmarks compared to the baseline model. Furthermore, on the Dyn-VQA benchmark, which requires multi-hop reasoning and web search capabilities, it demonstrates an improvement of 8.8%.

The main contributions of this study can be summarized as follows: (1) We propose the first million-scale multimodal agent tuning dataset, MMAT-1M, addressing a critical gap in the domain of multimodal agent tuning. (2) We establish a reflection mechanism that effectively mitigates logical errors in the reasoning process, significantly enhancing the model’s robustness. (3) We offer datasets in both

one-turn and iterative formats, providing flexibility to balance precision and efficiency in practical applications.

2. Related Work

LLM-based Agents LLM-based agents are primarily large models that harness the instruction-following capabilities of LLMs to develop advanced reasoning and tool-usage functionalities. Notable frameworks in this domain include HuggingGPT [33], GPT4Tools [44], VisualChatGPT [40], among others. ReAct [45], for instance, introduces a general paradigm that integrates CoT reasoning with action execution to address a broad spectrum of reasoning and decision-making challenges. Similarly, AssistGPT [13], proposes a “Learner” module that analyzes the prediction process and facilitates reflection, aligning with ReAct’s methodology. However, these approaches heavily rely on the instruction comprehension capabilities of LLMs, which restricts their effectiveness in handling longer or more complex reasoning tasks. Additionally, the high computational costs associated with invoking large models further raise the barrier to practical application.

Multimodal Agent Tuning. Agent tuning is a specialized subfield of language model fine-tuning, focused on enhancing the capabilities of LLMs in areas such as planning, reasoning, and tool usage. Among the earliest works in this domain are AgentTuning [49] and Fireact [3], which laid the foundation for subsequent advancements in agent tuning. Subsequently, many efforts are dedicated to advancing agent tuning [7, 34, 36, 47]. However, these methods primarily concentrate on optimizing LLMs, which, when applied in the multimodal domain, can only access information through multimodal tools. To address this limitation, several studies have explored multimodal agent tuning to improve reasoning and tool usage for multimodal challenges. For instance, LLaVA-Plus [23] represents the first attempt to train a multimodal assistant through visual instruction tuning, enabling it to learn tool usage effectively. Similarly, MLLM-Tool [37] is an agent system that integrates multimodal encoders with open-source LLMs to perceive and process instructions based on visual or audio inputs. Additionally, T3-Agent [14] generates a diverse range of multimodal tasks with detailed trajectories and leverages this data to fine-tune Vision-Language Models (VLMs) for enhanced tool utilization.

Multimodal Agent and CoT Dataset. To achieve strong performance in multimodal agent tuning, several datasets have been developed to optimize agents using diverse approaches. For instance, LLaVA-Plus transforms the LLaVA-158K dataset into a tool-use instruction format. Similarly, MLLM-Tool curates instruction-answer pairs encompassing 29 tasks sourced from HuggingFace. Meanwhile, T3-Agent introduces MM-Traj, a dataset comprising 20K trajectories, generated through a novel data collection

Statistics	Component	Number
Dataset Composition	Visual CoT [32]	434265
	LLaVA-CoT [43]	98561
	The Cauldron [19]	215680
	TabMWP [26]	23059
	Infoseek [6]	131400
Dialogue Turns	1 turn	846389
	2 turns	28646
	3+ turns	27930
Rationale Steps	2 turn	7909
	3 turns	763212
	4 turns	221440
	5+ turns	97702
Operator Calls	Image Caption	620644
	OVD	156237
	OCR	471866
	Face Detection	20077
	RAG	205682
Reflection Calls	General	46508
	Math	11139

Table 1. Key statistics of the MMAT-1M dataset.

pipeline. Moreover, some agents, such as OmniSearch [20], have designed the Dyn-VQA benchmark to evaluate capabilities in RAG and multi-hop reasoning tasks. In addition to these multimodal agent datasets, several multimodal CoT datasets share similar construction methodologies but lack explicit information on tool usage, such as LLaVA-CoT [42], Visual-CoT [32], and M³CoT [5].

3. MMAT-1M Dataset

In this section, we provide a comprehensive introduction to MMAT-1M, detailing its key components and methodologies. The discussion is structured into three parts: (1) an overview of the dataset, which outlines its scope, composition, and significance (Section 3.1); (2) the data engine, which describes the iterative framework for generating and refining high-quality trajectories (Section 3.2); and (3) the multimodal agent tuning method, which explains the approaches for enhancing reasoning and tool-usage capabilities (Section 3.3).

3.1. Overview of MMAT-1M

To build a diverse and comprehensive MMAT-1M dataset, we consolidate data from five distinct sources. These sources encompass a wide range of critical domains in multimodal tasks, including visual understanding, logical reasoning, mathematical computation, and knowledge retrieval. This integration ensures both the diversity and completeness of the dataset. The details of each dataset are as follows:

Visual CoT [32] encompasses a variety of tasks, such as document parsing, fine-grained understanding, general

Dataset	Size	APIs	Online Search	CoT	Reflection	Turns
LLaVA-Plus-v1 [23]	117K	✓	✗	✓	✗	multiple
Visual CoT [32]	438K	✗	✗	✓	✗	one
LLaVA-CoT [43]	100K	✗	✗	✓	✗	one
MM-Traj [14]	20K	✓	✓	✓	✗	one
MMAT-1M	1M	✓	✓	✓	✓	one&multiple

Table 2. Comparison of MMAT-1M with other training datasets.

visual question answering (VQA), chart analysis, and relational reasoning. Its primary objective is to strengthen models’ capabilities in focusing on localized visual regions and executing step-by-step reasoning processes. **LLaVA-CoT** [43] places a strong emphasis on complex reasoning and systematic thinking. It tackles a range of tasks, including general VQA, scientific reasoning, mathematical reasoning, and document understanding, aiming to enhance models’ hierarchical reasoning capabilities and improve their interpretability. **The Cauldron** [19] incorporates a wide array of multimodal data types, including interleaved text-image documents, text-image pairs, OCR-processed documents, and tables or charts. The diversity of its data sources and task designs plays a pivotal role in advancing models’ generalization capabilities, particularly in the integration of visual and linguistic information. **TabMWP** [26] focuses on mathematical reasoning tasks that integrate both textual and tabular data, seeking to improve models’ table parsing, numerical computation, and complex reasoning capabilities. **Infoseek** [6] is centered on visual information-seeking question answering, designed to assess and enhance the performance of multimodal models in knowledge-intensive visual question-answer tasks. These tasks demand fine-grained reasoning that extends beyond common sense and often relies on external knowledge bases for accurate responses.

The statistical information of the MMAT-1M dataset is shown in Table 1. The dataset comprises a total of 1,090,263 question-answer pairs and 902,965 dialogues, distributed across distinct subsets to ensure diversity in data sources. The second row of the table shows the number of dialogue turns in the original data, which shows that the one-turn dialogues represent the majority of samples, while the multi-turn dialogues are comparatively less frequent. In terms of reasoning complexity, the majority of data samples involve two-step and three-step reasoning processes, which serve as the foundational level of reasoning. In contrast, tasks requiring more intricate, multi-step reasoning constitute a smaller proportion, highlighting the dataset’s inclusion of both basic and advanced cognitive challenges. Meanwhile, among a wide range of operator calls, the invocation of Image Caption and OCR is relatively high, indicating the demand for basic information of images and text in the reasoning process. RAG and OVD also account for a notable proportion of operator invocations. Furthermore,

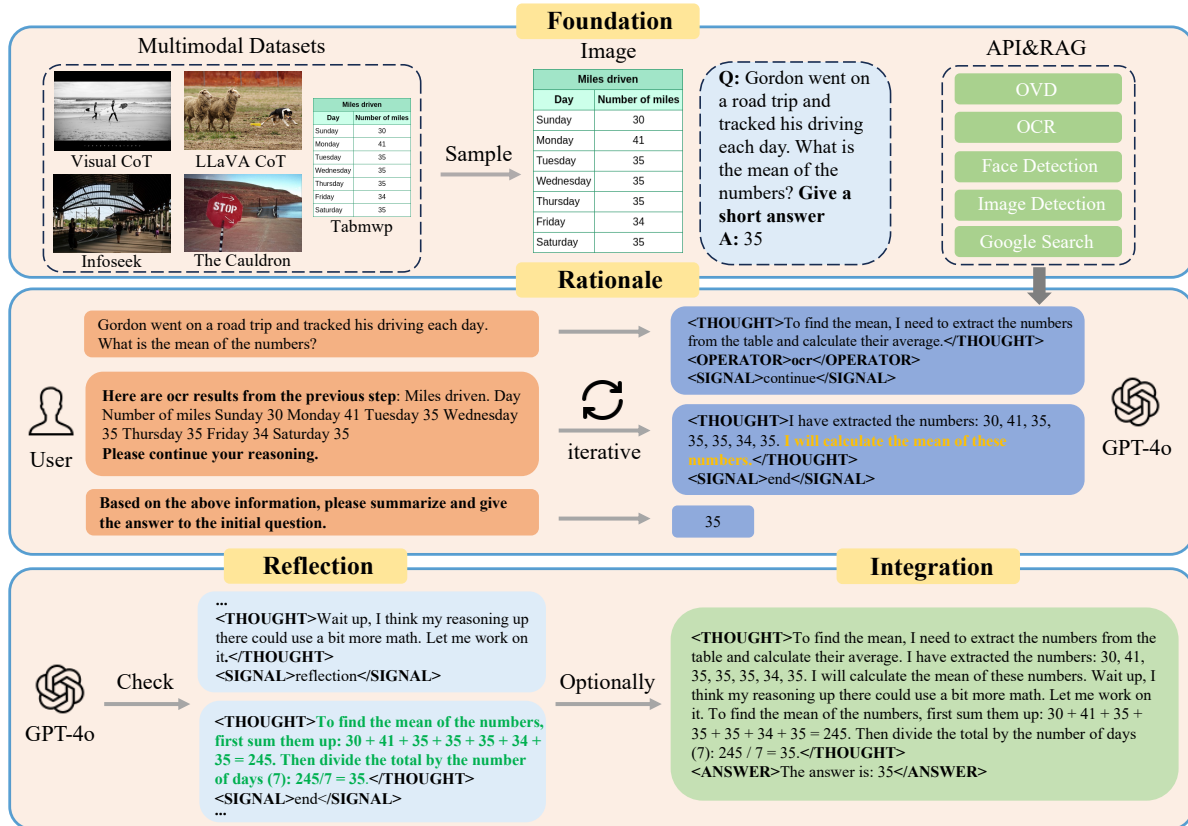


Figure 2. The data engine pipeline follows four stages: foundation, rationale generation, reflection, and trajectory integration. It generates datasets in two formats (RR and ORR) achieving a balance between precision and efficiency.

the reflection section encompasses both general reflection and mathematical reasoning reflection, comprising a total of approximately 57k data points. In summary, MMAT-1M is distinguished by its large-scale data volume, diverse task coverage, and hierarchical reasoning depth, collectively establishing a robust and flexible data foundation for advancing research in multimodal agent tuning.

We compare MMAT-1M with several similar agent tuning and CoT datasets, including LLaVA-Plus-v1 [23], Visual CoT [32], LLaVA-CoT [43] and MM-Traj [14], as shown in Table 2. It is evident that the scale of our dataset substantially exceeds that of comparable datasets. Furthermore, our dataset is equipped with API and RAG tool invocation capabilities, supports CoT reasoning and reflection, and encompasses both one-turn and multi-turn reasoning paradigms.

3.2. Data Engine

As shown in Figure 2, the data construction process is structured into four key stages: foundation, rationale generation, reflection, and integration of trajectories.

Foundation. As an illustrative example, we randomly select an image and its corresponding question-answer pair

from the original dataset. To ensure consistency in response styles across different datasets, we optimize the phrasing of the questions. For samples with shorter answers, we append a response style constraint at the end of the question, while keeping the original answer unchanged. Additionally, we prepare external tools for invocation, including Image Caption, OVD, OCR, Face Detection, and RAG. The Image Caption operator generates textual descriptions of images, extracting key visual information and expressing their semantics. Based on the CCoT [29], we use GPT-4o to construct a scene graph and generate image descriptions accordingly, enhancing semantic understanding and compositional reasoning capabilities. OVD leverages object information from the scene graph to identify and detect targets within an open vocabulary range, enabling the recognition of novel categories that extend beyond a predefined label set. This functionality is implemented using Grounding DINO [24]. OCR utilizes PaddleOCR [11] to recognize textual content within images. Face Detection, powered by deepface [30], accurately locates facial regions in images. Finally, for questions that require online search capabilities, we leverage GPT-4o to generate search queries, which are then used to invoke the Google API to retrieve the top-k

most relevant information.

Rationale. We employ an iterative diagram to generate rationales, where the annotation process is powered by GPT-4o, ensuring the stability and efficiency of reasoning. During inference, the model adaptively invokes multimodal operators RAG to maintain the completeness and interpretability of the reasoning chain. The reasoning process initiates with problem analysis, where the model selects appropriate operators based on task requirements. If holistic semantic understanding is required, the Image Caption operator is invoked to extract a scene graph and generate an image description. For tasks demanding object-level information, the OVD operator is utilized to identify objects within an open-vocabulary range. Similarly, the OCR operator and Face Detection operator are employed for text recognition and facial analysis, respectively. When operator outputs are insufficient to support inference, the model formulates RAG queries to retrieve and integrate external knowledge. Each reasoning step is meticulously recorded in a structured STRING format, capturing inference thoughts, operator invocations, retrieval requests, and subsequent actions. This adaptive multi-turn reasoning mechanism ensures the logical coherence of the reasoning chain, ultimately producing accurate, interpretable, and well-documented rationales.

Reflection. In our observations, the rationales generated through the process mentioned above exhibit two notable issues. The first is incompleteness in the reasoning process, particularly evident in the derivation of mathematical problems. This occurs when certain steps are omitted, making it challenging to arrive at the final answer. The second issue is reasoning cheating behavior, where the rationale’s thought process does not logically lead to the final answer, but GPT-4o forcibly aligns the reasoning with the answer during label generation, creating an illusion of correctness. To address these issues, we introduce reflective steps aimed at enhancing the model’s error-correction capabilities during training and ensuring the reasoning process remains logically sound. Specifically, for the first issue, GPT-4o is tasked with identifying whether “step skipping” behavior exists in the reasoning process. If such behavior is detected, missing steps are supplemented to complete the derivation. For the second issue, we employ GPT-4o to re-evaluate whether the rationale’s thought process aligns with the final answer. If a mismatch is identified, a reflective process is implemented to make the rationale aware of the cheating behavior and correct it accordingly.

Integration. The dataset generated through the approach above adopts a multi-turn Rationale and Reflection (RR) format, which may be impractical for real-world applications requiring time-sensitive responses. Inspired by the LUMOS [47] model, we aim to create a dataset where the model can deliberate and produce the final answer in one turn. However, due to the constraints of the one-turn for-

mat, we cannot dynamically incorporate the results of external operators during the output phase. To address this, we integrate the results of all operators (excluding RAG) into the input stage, clearly demarcated by brackets. At the output stage, we consolidate multiple trajectories from the multi-turn dialogue into a One-turn Rationale and Reflection (ORR) format. Our findings indicate that ORR not only retains the ability to perform reasoning and integrate external tool results but also significantly improves inference speed, making it more suitable for time-critical applications.

To assess potential GPT-4o hallucinations, we evaluated all MMAT-1M samples on coherence, relevance, accuracy, completeness, and image-text alignment, with over 89% demonstrating high-quality reasoning. Evaluation criteria are detailed in the supplementary material.

3.3. Multimodal Agent Tuning

Given a training sample: $\{\{q_1, r_1\}, \dots, \{q_i, r_i\}, \dots, \{q_n, A\}\}$, where q_i is i -th question, r_i indicates the rationale, and A signifies the final answer. We select several open-source multimodal models and employ supervised fine-tuning (SFT) training schemes on these models.

SFT. We opt for low-rank adaptation (LoRA) [16], compared to full parameters fine-tuning, which not only retains the majority of the baseline model’s knowledge but also save memory and computational space efficiently. The loss function of it is designed as follows:

$$L = L_{\text{original}} + \lambda \sum_i \|\Delta\theta_i\|_F^2, \quad (1)$$

where L_{original} is the original loss function, $\Delta\theta_i$ indicates the update of the i -th weight matrix, λ is the regularization parameter, and $\|\cdot\|_F$ denotes the Frobenius norm.

4. Experiments

We conduct extensive experiments across multiple benchmarks to evaluate the effectiveness of our approach. Section 4.1 details the implementation settings. In Section 4.2, we compare our method, which fine-tunes MLLMs with One-turn Rationale and Reflection (ORR) and Rationale and Reflection (RR) strategies on the MMAT-1M dataset, against baselines. The evaluation spans eight benchmarks, covering general and reasoning tasks, along with one benchmark for external knowledge retrieval. Section 4.3 presents ablation studies and analyzes inference efficiency. Finally, Section 4.4 provides qualitative results for further insights into our method.

4.1. Implementation Details

In this section, we integrate MMAT-1M with various MLLMs to showcase the broad applicability of our approach. We investigate two reasoning strategies, ORR and

Model	Method	Average	MMStar	MMMU	MathVista	MathVision	AI2D	OCRBench	RealWorldQA	HallusionBench
GPT-4o [17]	/	65.6	65.1	70.7	60.0	30.4	84.9	806	76.5	56.2
Llama-3.2-11B-Vision-Instruct [28]	Baseline	52.2	47.7	50.3	48.0	16.4	77.1	756	63.4	39.4
	ORR	54.6	50.7	47.8	50.1	17.7	78.9	806	66.7	44.4
	RR	55.3	51.4	51.0	49.1	16.8	77.9	784	69.3	48.3
MiniCPM-V-2.6 [46]	Baseline	58.0	56.5	47.1	60.3	22.4	81.5	843	65.0	47.1
	ORR	58.8	56.9	47.9	60.6	23.4	81.7	848	66.6	48.8
	RR	59.9	58.5	49.2	61.9	25.3	82.0	840	68.0	50.0
InternVL2.5-2B [8]	Baseline	52.7	53.6	43.2	50.1	16.1	75.1	804	60.5	42.6
	ORR	54.4	55.4	44.7	50.1	14.1	77.5	819	69.5	42.4
	RR	54.7	54.9	44.4	52.6	16.5	77.2	799	68.0	43.8
InternVL2.5-4B [8]	Baseline	58.4	58.6	51.8	60.8	21.7	81.2	823	64.6	46.5
	ORR	59.5	59.2	50.7	61.4	19.7	81.4	824	69.2	51.9
	RR	60.6	60.9	53.1	62.0	22.4	82.7	805	72.2	50.7
InternVL2.5-8B [8]	Baseline	60.7	62.4	53.1	64.5	20.1	84.1	819	69.4	49.8
	ORR	62.4	64.8	55.4	63.8	20.8	83.5	849	73.0	53.3
	RR	63.4	65.3	57.3	64.8	21.7	84.2	839	74.4	55.8

Table 3. Performance comparison of MLLMs with Baseline, ORR (One-turn Rationale and Reflection), and RR (Rationale and Reflection) across eight benchmarks. Models trained on MMAT-1M with ORR and RR achieve overall gains, enhancing multimodal capabilities.

Model	Query	Golden Query
GPT-4o [17]	52.0	61.5
OmniSearch (GPT-4V) [20]	50.0	/
Llama-3.2-11B-Vision-Instruct [28]	29.4	34.6
Llama-3.2-11B-Vision-Instruct-RR	38.0	45.1
MiniCPM-V-2.6 [46]	32.7	39.2
MiniCPM-V-2.6-RR	35.9	44.4
InternVL2.5-2B [8]	19.3	26.0
InternVL2.5-2B-RR	30.9	38.8
InternVL2.5-4B [8]	23.3	31.1
InternVL2.5-4B-RR	35.4	42.1
InternVL2.5-8B [8]	27.0	35.2
InternVL2.5-8B-RR	36.8	44.0

Table 4. Results on the RAG Benchmark Dyn-VQA. RR strategy significantly boosts performance across model scales, enhancing multi-hop reasoning and retrieval.

RR, which guide multimodal models toward structured and interpretable reasoning. ORR consolidates all reasoning steps into a single query, enabling efficient inference while maintaining strong accuracy. In contrast, RR follows a multi-step reasoning process, dynamically selecting operators and retrieving external knowledge when needed. For reasoning scenarios that require external knowledge injection, we employ Google Search to retrieve relevant information. Each query returns up to three results (top-k=3), providing the model with necessary contextual knowledge while maintaining efficiency.

We apply these strategies to open-source multimodal models, including Llama-3.2-11B-Vision-Instruct [28], MiniCPM-V-2.6 [46], and the InternVL2.5 series [8], which includes InternVL2.5-2B, InternVL2.5-4B, and InternVL2.5-8B. Each model is separately fine-tuned with ORR and RR on the MMAT-1M dataset, which consists of 1,090,263 question-answer pairs, for one epoch with a

learning rate of 4e-5. Detailed training parameters are provided in the supplementary material.

4.2. Main Results on Benchmark

Setup. We conduct a comprehensive evaluation of our method using eight widely adopted and challenging benchmarks: MMStar [4], MMMU [48], MathVista [27], MathVision [38], AI2D [18], OCRBench [25], RealWorldQA [41], and HallusionBench [15]. Specifically, MMStar and MMMU primarily assess multimodal reasoning and question-answering capabilities, while MathVista and MathVision focus on mathematical and visual reasoning skills. AI2D examines the comprehension of scientific diagrams, and OCRBench evaluates textual information extraction from documents. RealWorldQA targets spatial reasoning in real-world scenarios, whereas HallusionBench gauges susceptibility to language hallucinations and visual illusions. For MathVista and MathVision, we adopt the testmini set. To ensure fairness and reproducibility, all evaluations are conducted using VLMEvalKit [12], an open-source toolkit specifically designed for large vision-language models. Beyond these benchmarks, we further evaluate the RAG capabilities of the models with the Dyn-VQA dataset proposed in OmniSearch [20]. Dyn-VQA encompasses dynamic, multimodal, multi-hop reasoning tasks, offering a comprehensive assessment of how effectively models plan retrieval strategies and integrate relevant information.

Main Results. Table 3 presents experimental results on multiple benchmarks that evaluate the performance of various multimodal large models trained on MMAT-1M using ORR and RR. The findings demonstrate that both methods effectively enhance model performance across different parameter scales.

Training with our ORR on MMAT-1M improves the av-

Model	API	RAG	Average	MMStar	MMMU	MathVista	MathVision	AI2D	OCRBench	RealWorldQA	HallusionBench	Dyn-VQA
Baseline	×	×	57.9	62.4	53.1	65.1	20.1	84.1	819	69.4	49.8	35.2
Baseline-RR	✓	×	59.8	65.0	56.2	64.2	20.4	84.1	839	74.3	55.0	35.4
Baseline-RR	×	✓	57.3	60.1	52.6	61.1	21.0	81.8	797	67.8	48.0	43.4
Baseline-RR (w/o SFT)	✓	✓	55.0	60.6	49.8	60.9	15.1	82.8	825	68.9	43.2	31.5
Baseline-R	✓	✓	60.2	65.0	54.5	63.9	20.5	84.6	826	72.7	54.8	42.9
Baseline-ORR	✓	×	59.6	64.8	55.4	63.8	20.8	83.5	849	73.0	53.3	36.6
Baseline-RR	✓	✓	61.3	65.3	57.3	64.8	21.7	84.2	839	74.4	55.8	44.0

Table 5. Ablation study evaluating the impact of SFT, API integration, structured reflection, and RAG-based retrieval on multimodal reasoning performance. Results highlight the complementary benefits of fine-tuning, explicit rationale generation, and external knowledge integration in enhancing multimodal reasoning performance.

erage score of InternVL2.5-8B from 60.7 to 62.4 compared to the baseline, while our RR strategy further boosts it to 63.4. Notably, RR consistently outperforms the baseline and achieves competitive results against GPT-4o. Specifically, InternVL2.5-8B with RR surpasses GPT-4o on MMStar (65.3 vs. 65.1) and MathVista (64.8 vs. 60.0), demonstrating superior multimodal reasoning and mathematical-visual understanding. It also outperforms GPT-4o on OCRBench (839 vs. 806), reflecting stronger textual information extraction. Additionally, it performs on par with GPT-4o on AI2D (84.2 vs. 84.9) and HallusionBench (55.8 vs. 56.2), indicating robust comprehension of scientific diagrams and resilience to multimodal hallucinations.

Compared with baseline models such as InternVL2.5-8B, MiniCPM-V-2.6, and Llama-3.2-11B-Vision-Instruct, our ORR and RR particularly RR, have demonstrated generally similar optimization effects across various test sets. Our RR on MiniCPM-V-2.6 achieves a gain in average from 58.0 to 59.9, a 3.3% relative increase, while on Llama-3.2-11B-Vision-Instruct achieves a gain from 52.2 to 55.3, a relative improvement of 5.9%. This indicates that our methods have broad applicability across different model series. Similarly, our ORR and RR consistently deliver strong performance across the InternVL2.5 series, including the 2B, 4B, and 8B parameter variants, demonstrating robust scalability and wide-ranging applicability of our methodology.

In OCRBench, InternVL2.5-2B’s ORR strategy outperforms the baseline (804 to 819), while RR drops to 799, a trend also seen in InternVL2.5-4B and 8B. The reason for this phenomenon is that, although RR exhibits specific error-correction capabilities, the OCR misrecognition negatively impacts the final results. In contrast, ORR utilizes image captioning to mitigate OCR errors, demonstrating superior performance in OCRBench.

The comprehensive results confirm that training on MMAT-1M with our ORR and RR leads to significant improvements, particularly with RR, in tasks requiring comprehensive reasoning, mathematical computation, and cross-modal information fusion. This establishes MMAT-1M as a valuable benchmark for advancing the reasoning capabilities of vision-language models.

Results on RAG Benchmark. The evaluation results of Dyn-VQA [20] are shown in Table 4, based on the latest version. Query refers to the input content used by the model for information retrieval, while Golden Query denotes an optimized prompt focused on the final retrieval step to maximize answer accuracy. To align with Dyn-VQA, we adopt the same evaluation metric, F1-Recall, which measures the overlap between the model-generated response and the ground truth. Results demonstrate that our ORR and RR consistently enhance multi-hop reasoning and retrieval performance. Specifically, the RR improves Llama-3.2-11B-Vision-Instruct by 29.3% relative to its original performance (from 29.4 to 38.0) in Query and by 30.3% relative to its original performance (from 34.6 to 45.1) in Golden Query, while MiniCPM-V-2.6 shows improvements of 9.8% and 13.3%, respectively. The InternVL2.5 series models similarly benefit, with relative gains ranging from 31.9% to 60.1%, underscoring the effectiveness of our methods across complex, knowledge-intensive tasks.

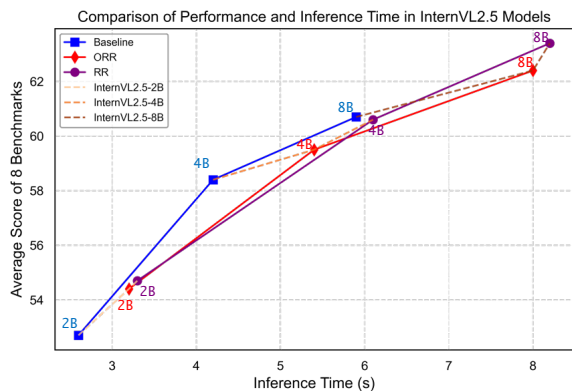


Figure 3. Comparison of inference efficiency and performance gains of ORR and RR across different InternVL2.5 model scales.

4.3. Further Analysis

Ablation Study. Table 5 presents an ablation study on the effects of SFT, API integration, structured reflection, and RAG on multimodal reasoning performance. The baseline model, without external resources, achieves an aver-

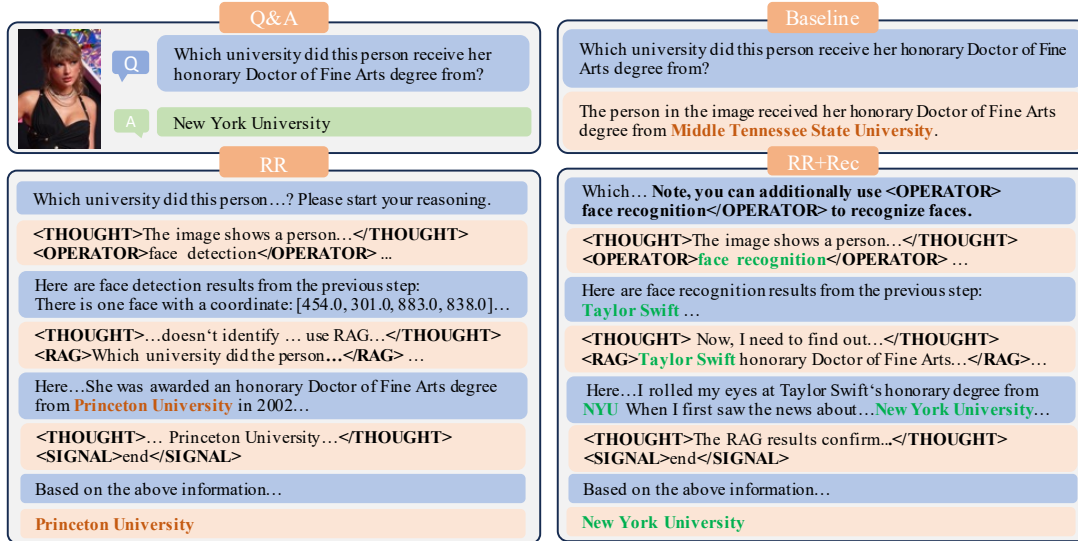


Figure 4. The zero-shot capability of invoking a celebrity recognition operator of InternVL2.5-8B-RR.

age score of 57.9. In the RR setting, Baseline-RR achieves the highest score of 61.3 with both API and RAG. Removing API reduces performance to 57.3, while removing RAG lowers it to 59.8. Without SFT, performance declines further to 55.0. Additionally, Baseline-R, which retains rationale but omits reflection, scores 60.2, suggesting that reflection enhances reasoning ability. In the ORR setting, performance declines to 59.6, primarily because the ORR format does not incorporate RAG information, resulting in a performance drop on the Dyn-VQA benchmark. On other benchmarks, however, its performance remains comparable to that of the RR format. These results confirm that SFT is crucial for instruction adherence, while structured reflection and external knowledge integration further improve multimodal reasoning.

Performance Efficiency Tradeoff between ORR and RR.

Figure 3 compares the inference efficiency and performance gains of the ORR and RR methods across different InternVL2.5 model scales. Although both ORR and RR consistently enhance multimodal reasoning performance, their inference times notably increase relative to the baseline. ORR introduces a moderate inference overhead due to its one-turn structured reasoning approach, while RR, involving multi-turn adaptive reasoning steps, incurs a slightly higher computational cost. However, RR achieves greater performance improvements compared to ORR, demonstrating a beneficial tradeoff between computational efficiency and reasoning accuracy.

4.4. Qualitative Results

While the experiments, as mentioned above, have demonstrated the benefits of invoking external tools for the model, the capabilities of a fixed set of tools are inherently limited.

For instance, MMAT-1M’s lack of a celebrity recognition operator hinders the fine-tuned model from achieving correct results in cases requiring celebrity identification. To address this, we conduct an experiment to verify whether the fine-tuned model can invoke operators it has not been explicitly trained on. As shown in Figure 4, we test a visual question with the InternVL2.5-8B model. Initially, the baseline model provides an incorrect answer. As anticipated, the model fine-tuned on MMAT-1M, failing to recognize the person, also returns a wrong answer due to unsuccessful web search results. To address this limitation, we instruct the fine-tuned model to invoke a celebrity recognition operator, which successfully identifies the correct answer. This experiment demonstrates that the model fine-tuned on our dataset exhibits a certain level of zero-shot capability for invoking unseen tools. However, its performance remains inferior to that achieved through explicit fine-tuning.

5. Conclusion

The introduction of MMAT-1M represents a significant advancement in multimodal agent tuning, offering a diverse and flexible dataset for enhancing CoT reasoning and tool usage in MLLMs. By addressing key limitations of existing multimodal agent tuning datasets, such as homogeneity, lack of reflection, and inflexible tool usage, it provides a comprehensive solution that aligns with the demands of real-world applications. While the dataset demonstrates robust performance on current multimodal benchmarks, further research is essential to evaluate its adaptability to a broader array of MLLMs and more intricate real-world scenarios.

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(8), 2023. 1
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1
- [3] Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023. 1, 2
- [4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 6
- [5] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M³ cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024. 3
- [6] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. 3
- [7] Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. Agentflan: Designing data and methods of effective agent tuning for large language models. *arXiv preprint arXiv:2403.12881*, 2024. 1, 2
- [8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 6
- [9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.
- [10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1
- [11] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020. 4
- [12] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 6
- [13] Difei Gao, Lei Ji, Luwei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023. 1, 2
- [14] Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 4
- [15] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 6
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5
- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6
- [18] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 6
- [19] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 3
- [20] Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Pengjun Xie, Philip S Yu, et al. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*, 2024. 3, 6, 7
- [21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1
- [22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1
- [23] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multi-

- modal agents. In *European Conference on Computer Vision*, pages 126–142. Springer, 2024. 2, 3, 4
- [24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 4
- [25] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv e-prints*, pages arXiv–2305, 2023. 6
- [26] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 3
- [27] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 6
- [28] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*. Retrieved December, 20:2024, 2024. 6
- [29] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 4
- [30] Sefik Serengil and Alper Özpınar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilişim Teknolojileri Dergisi*, 17(2):95–107, 2024. 4
- [31] Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782, 2024. 1
- [32] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 3, 4
- [33] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023. 1, 2
- [34] Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization for llm agents. *arXiv preprint arXiv:2403.02502*, 2024. 2
- [35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [36] Boshi Wang, Hao Fang, Jason Eisner, Benjamin Van Durme, and Yu Su. Llms in the imaginary: tool learning through simulated trial and error. *arXiv preprint arXiv:2403.04746*, 2024. 1, 2
- [37] Chenyu Wang, Weixin Luo, Qianyu Chen, Haonan Mai, Jindi Guo, Sixun Dong, XM Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. Mllm-tool: A multimodal large language model for tool agent learning. *arXiv preprint arXiv:2401.10727*, 4, 2024. 2
- [38] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2025. 6
- [39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1
- [40] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 1, 2
- [41] X.AI. Grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v>, 2024. 6
- [42] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 3
- [43] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024. 3, 4
- [44] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36:71995–72007, 2023. 1, 2
- [45] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2
- [46] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6
- [47] Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, and Bill Yuchen Lin. Agent lumos: Unified and modular training for open-source language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12380–12403, 2024. 1, 2, 5
- [48] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming

Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. [6](#)

- [49] Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. Agenttuning: Enabling generalized agent abilities for llms. *arXiv preprint arXiv:2310.12823*, 2023. [1](#), [2](#)