

A Survey on Agentic Multimodal Large Language Models

Huanjin Yao[†], Ruifei Zhang[†], Jiaxing Huang[✉], Jingyi Zhang, Yibo Wang, Bo Fang, Ruolin Zhu, Yongcheng Jing, Shunyu Liu, Guanbin Li, Dacheng Tao

Abstract—With the recent emergence of revolutionary autonomous agentic systems, research community is witnessing a significant shift from traditional static, passive, and domain-specific AI agents toward more dynamic, proactive, and generalizable agentic AI. Motivated by the growing interest in agentic AI and its potential trajectory toward AGI, we present a comprehensive survey on Agentic Multimodal Large Language Models (Agentic MLLMs). In this survey, we explore the emerging paradigm of agentic MLLMs, delineating their conceptual foundations and distinguishing characteristics from conventional MLLM-based agents. We establish a conceptual framework that organizes agentic MLLMs along three fundamental dimensions: (i) **Agentic internal intelligence** functions as the system’s commander, enabling accurate long-horizon planning through reasoning, reflection, and memory; (ii) **Agentic external tool invocation**, whereby models proactively use various external tools to extend their problem-solving capabilities beyond their intrinsic knowledge; and (iii) **Agentic environment interaction** further situates models within virtual or physical environments, allowing them to take actions, adapt strategies, and sustain goal-directed behavior in dynamic real-world scenarios. To further accelerate research in this area for the community, we compile open-source training frameworks, training and evaluation datasets for developing agentic MLLMs. Finally, we review the downstream applications of agentic MLLMs and outline future research directions for this rapidly evolving field. To continuously track developments in this rapidly evolving field, we will also actively update a public repository at <https://github.com/HJYao00/Awesome-Agentic-MLLMs>.

Index Terms—Agentic MLLMs, Reinforcement Learning, Reasoning, Reflection, Memory, Search, Code, Thinking with images

1 INTRODUCTION

MULTI-MODAL Large Language Models (MLLMs) have achieved remarkable progress in recent years, enabling AI systems to perceive, understand, reason, and generate across diverse modalities [1, 2, 3, 4, 5, 6, 7, 8]. With strong instruction-following ability and cross-modal generalization, MLLMs are capable of tackling a wide spectrum of tasks, making them increasingly valuable in both general applications and professional contexts [9, 10, 11, 12, 13, 14]. However, most traditional MLLMs still operate under a query-response paradigm, where static inputs produce single outputs. This paradigm is often inadequate for complex, dynamic real-world tasks, which require three essential capabilities: internal intelligence (e.g., reasoning [15, 16, 17, 18], reflection [19, 20], and memory [21, 22]), external tool invocation (e.g., information searching [23, 24], code execution [25, 26], and visual processing [27, 28, 29]), and environment interaction (e.g., virtual embodiment [30, 31] and physical embodiment [32, 33]).

To extend the capabilities of MLLMs beyond static query-response interactions, MLLM agents [34, 35] have attracted increasing attention, which embeds MLLMs within structured workflows, enabling task decomposition, scenario-specific reasoning, and integration of external

tools [36, 37, 38, 39, 40, 41]. Despite their effectiveness, existing MLLM agents still suffer from several constraints: 1) Static workflow: they rely heavily on pre-defined and handcrafted workflows that are inflexible and cannot adapt to novel or dynamic situations; 2) Passive execution: they typically respond passively to instructions, without genuine intelligence to initiate plans, invoke tools, or proactively engage with environments; 3) Domain-specific application: most MLLM agents are tailored for a single task or domain, resulting in poor generalization and limited scalability across diverse domains or tasks.

Recent advances in reasoning-enhanced MLLMs [47, 61, 182, 183] and reinforcement learning (RL) [184, 185, 186, 187] have driven a paradigm shift from workflow-bound MLLM agents toward agentic MLLMs. Unlike traditional agents, agentic MLLMs [24, 160, 161, 163, 188, 189] are framed as autonomous decision-makers, which possess built-in agentic capabilities, i.e., the autonomy to reason, reflect, memory, use tools, and interact with environments. To this end, agentic MLLMs offer several key advantages: (1) First, agentic MLLMs can dynamically adjust their strategies and workflows based on previous planning, current state, and anticipated environmental interactions rather than relying on static, pre-defined and handcrafted procedures. (2) Second, agentic MLLMs plan and execute actions proactively, autonomously initiating plans, invoking tools when needed, and reflecting on intermediate outcomes to refine subsequent steps. (3) Third, agentic MLLMs can operate across diverse tasks and environments, enabling general-purpose modeling and learning, instead of being restricted to narrow, domain-specific applications. This transition marks not only

- Huanjin Yao, Jiaxing Huang, Jingyi Zhang, Yibo Wang, Yongcheng Jing, Shunyu Liu, Dacheng Tao are with the Nanyang Technological University, Singapore.
- Ruifei Zhang is with the Chinese University of Hong Kong, Shenzhen, China, and also with the Shenzhen Research Institute of Big Data, China.
- Guanbin Li is with the Sun Yat-sen University, China.
- Bo Fang is with the City University of Hong Kong, China.
- Ruolin Zhu is with the Communication University of China, China.
- [†] denotes equal contribution; [✉] denotes corresponding author.

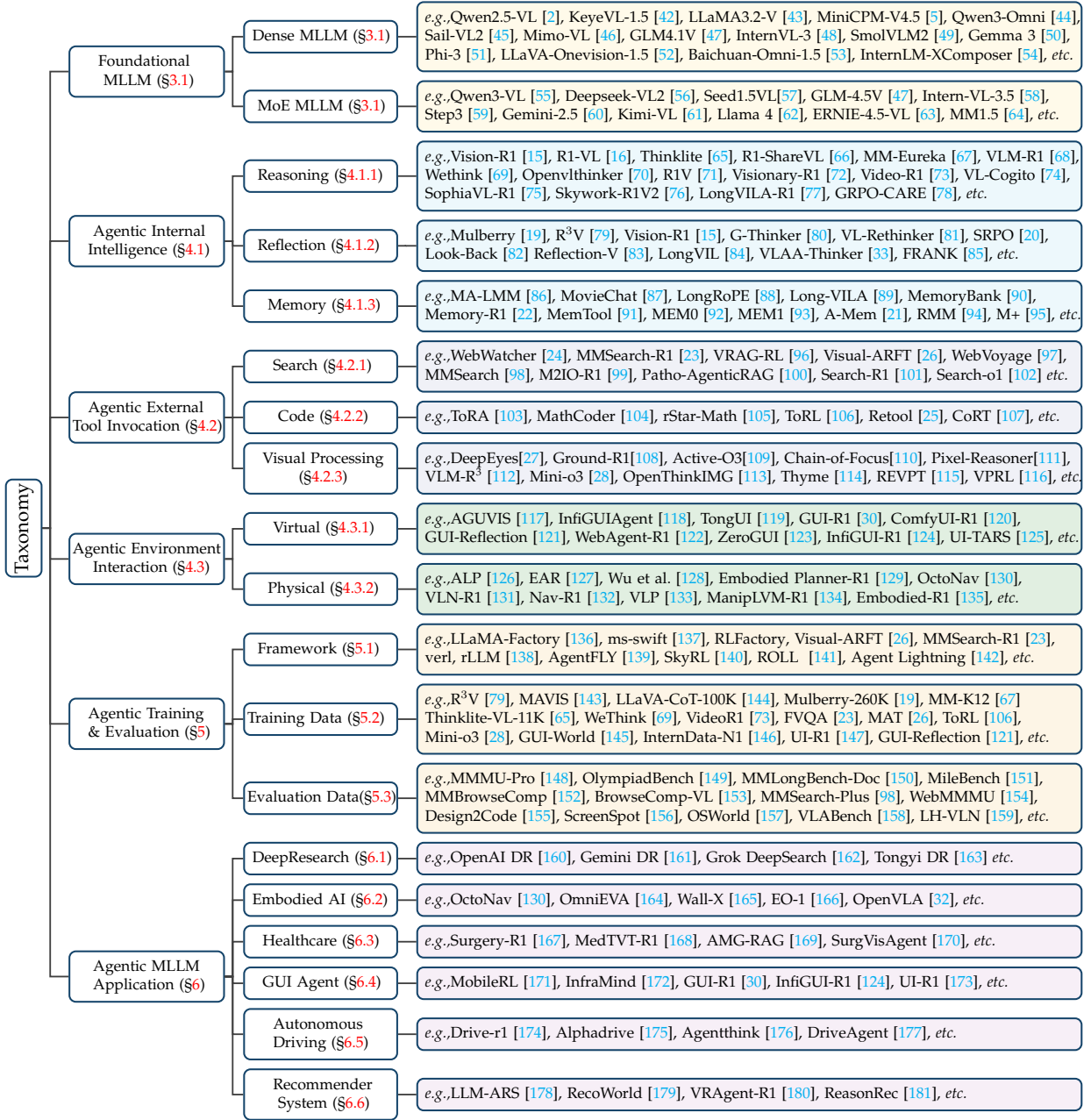


Fig. 1: The primary organizational structure of the survey and key works illustrating progress in each direction.

stronger planning capabilities, but also genuine intelligence: the ability to generate plans adaptively, invoke tools proactively, and engage effectively with dynamic environments.

Despite the growing attention on advancing agentic MLLMs, the research community still lacks a comprehensive survey that can help organize current progress, identify key challenges, and highlight promising directions in this rapidly evolving field. To fill this gap, we present a systematic review of agentic MLLMs over three major components including agentic internal intelligence, agentic external tool invocation, and agentic environment interaction. We conduct the survey from different perspectives including discussion, foundations, technical approaches, training & evaluation resources, and future research directions. We expect this survey to provide a thorough overview of current achievements and to outline the pathways for further

progress in this rapidly evolving and promising area.

In summary, the main contributions of this work are threefold. 1) it presents a systematic review of the development of agentic MLLMs, categorizing existing studies according to different tasks. To the best of our knowledge, this is the first survey in this field, offering an overarching view and thorough classification. 2) it studies the up-to-date progress of agentic MLLMs, including methodological advances as well as training and evaluation resources, with corresponding links provided for ease of reference. 3) it shares several research challenges and potential research directions that could be pursued in agentic MLLMs.

To this end, our survey is organized according to the taxonomy illustrated in Figure 1. The rest of this survey is organized as follows. Section 2 presents the discussion of MLLM agents and agentic MLLMs. We then introduce

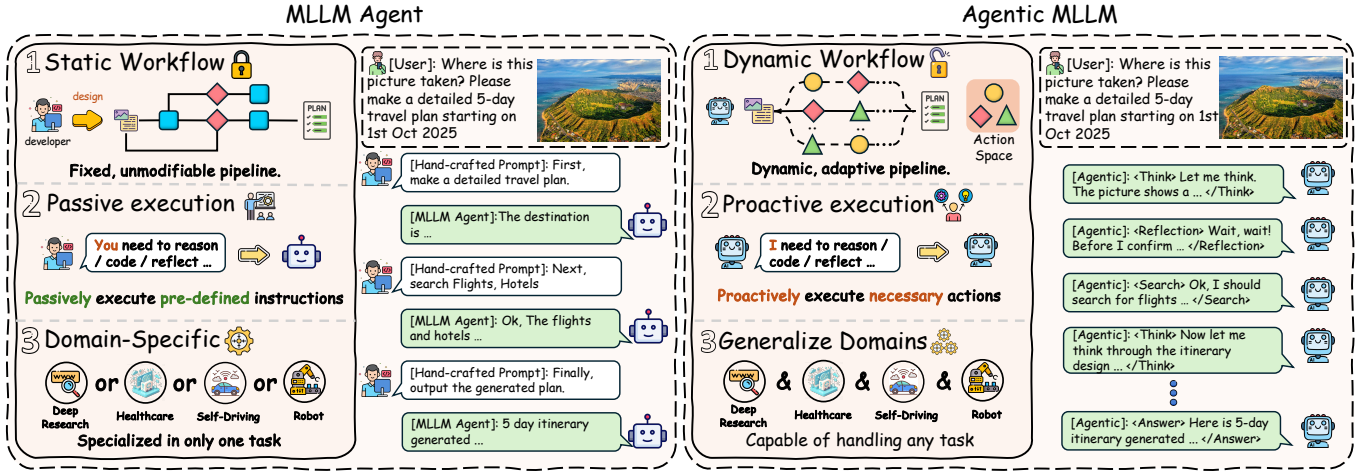


Fig. 2: The key differences between Agentic MLLMs and MLLM Agents lie in three defining characteristics of Agentic MLLMs: a dynamic and adaptive workflow, proactive execution of actions, and strong generalization across domains.

the foundational concepts of agentic MLLMs in Section 3, encompassing foundational MLLMs, agentic action space, agentic MLLM training and evaluation. Section ?? reviews and categorizes existing agentic MLLM studies, including agentic internal intelligence, agentic external tool invocation, and agentic environment interaction. Section 5 presents the widely-used training frameworks, training and evaluation datasets for agentic MLLMs. Section 6 introduces the applications of agentic MLLMs, such as DeepResearch, Embodied AI, Healthcare, GUI Agents, Autonomous Driving, and Recommender System. Finally, we share several promising agentic MLLMs research directions in Section 7.

2 DISCUSSION OF MLLM AGENT AND AGENTIC MLLM

This section formalizes the key distinctions between agentic MLLMs and conventional MLLM agents, emphasizing the dynamic workflows, proactive action execution, and cross-domain generalization capabilities of agentic MLLMs, as illustrated in Figure 2.

2.1 Overview of MLLM Agents

MLLM agents [34, 97, 190, 191, 192] are typically defined by a static workflow that is meticulously pre-designed and implemented by developers, adhering to a divide-and-conquer principle [193]. In this paradigm, a complex task is decomposed in a flowchart-like structure into a series of smaller subtasks, with the MLLM assigned different roles through carefully crafted prompts at each stage. Then, these role-specific instances of the MLLM execute their respective instructions within an orchestrated workflow, where intermediate outputs are cascaded downstream to subsequent stages. Ultimately, the process yields a complete solution in a modular manner, which can be formalized as follows:

$$\text{Agent}_{\text{MLLM}} = f_T \circ f_{T-1} \circ \dots \circ f_1(x_1) \quad (1)$$

$$f_i(x_i) = \text{MLLM}(p_i, x_i), x_{i+1} = f_i(x_i). \quad (2)$$

where p_i represents the manually crafted prompt at stage i , f_i denotes the responses of the MLLM conditioned on prompt p_i , and x_{i+1} is the sequential multimodal input passed forward from the previous stage. After all subtasks are completed in sequence, the overall process of $\text{Agent}_{\text{MLLM}}$ ultimately produces the final output.

Overall, MLLM agents position the MLLM as a task executor capable of accomplishing complex objectives through systematic decomposition into subtasks. However, their intrinsic design is bound to a static and fixed workflow, where roles are assigned exclusively through predefined prompts. This constraint results in static planning, passive action execution, and domain-specific limitations, as illustrated in Figure 2, hindering adaptability and generalizability.

2.2 Overview of Agentic MLLMs

In contrast, agentic MLLMs treat task-solving as an autonomous decision-making process, in which the model independently selects actions at each step in response to contextual features and evolving environmental states. As illustrated in Figure 2, we highlight three fundamental distinctions between MLLM agents and agentic MLLMs, which are elaborated in the following subsections.

2.2.1 Dynamic Workflow

As shown in Figure 2, traditional MLLM agents rely on a static and unmodifiable pipeline pre-designed by developers to solve the problem. In contrast, agentic MLLMs dynamically select appropriate strategies in response to the evolving state, enabling an adaptive problem-solving process and breaking free from fixed execution patterns. This dynamic workflow and its underlying state transitions can be represented at each step as:

$$s_{t+1} = \delta(s_t, a_t), \quad (3)$$

where s_t denotes the current state, a_t is the action chosen by the MLLMs, and δ represents the state transition function.

2.2.2 Proactive Action Execution

As illustrated in Figure 2, conventional MLLM agents passively execute actions at each stage according to pre-defined instructions designed by developers. In contrast, agentic MLLMs adopt a proactive paradigm in which actions are autonomously selected at every step based on the current state. This shift moves the model from simply following instructions to actively planning about “what action should be taken next,” thereby substantially improving its capacity for context-sensitive decision-making. Formally, proactive action execution can be expressed as:

$$a_t \sim \pi(a | s_t), \quad (4)$$

where a_t denotes the action chosen under the current state s_t according to the policy π .

2.2.3 Generalization Across Domains

As illustrated in Figure 2, traditional MLLM agents require developers to design bespoke pipelines and prompts for each task, rendering them domain-specific and limiting their ability to generalize to new scenarios. In contrast, agentic MLLMs can adapt their workflows across evolving environments by adaptively planning and executing the actions required. This flexibility enables them to operate in diverse contexts and to effectively solve tasks spanning multiple domains. Formally, such Generalization can be formulated as a policy optimization objective that maximizes the expected cumulative reward:

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{(x) \sim \mathcal{D}} \left[\sum_{t=0}^T \gamma^t r(s_t, a_t; x) \right], \quad (5)$$

where \mathcal{D} denotes the distribution of tasks and environments, s_t is the state at step t , a_t is the action sampled from policy π , $r(\cdot)$ is the reward function that drives generalization across domains, and γ is the discount factor controlling the relative importance of long-term versus short-term rewards.

In summary, agentic MLLMs reconceptualize task-solving within the formalism of an action-oriented markov decision process. Rather than relying on static, hand-crafted pipelines, they are modeled as adaptive policies that interact with action space and environment, continually updating internal states and proactively making context-sensitive decisions. This formulation highlights their ability to autonomously plan, act, and generalize across diverse tasks and domains.

$$\text{Agentic}_{\text{MLLM}} = \pi^*(x, \mathcal{A}, \mathcal{E}), \quad (6)$$

where x denotes the input, \mathcal{A} the action space, and \mathcal{E} the environment. Here, π^* represents the optimal policy that governs adaptive decision-making across states, actions, and environmental dynamics.

3 FOUNDATIONS OF AGENTIC MLLMS

In this section, we introduce the preliminaries of agentic MLLMs covering: (1) agentic foundational MLLMs, which serve as the base models for agentic systems; (2) agentic action space, which defines how actions are formally specified and subsequently executed by the model; (3) agentic continual pre-training, which equips MLLMs with broader

agentic general knowledge; (4) agentic supervised fine-tuning, which uses curated high-quality multi-turn trajectories to provide a cold start for RL; (5) agentic reinforcement learning, which incentivizes agentic behavior through exploration and feedback; (6) agentic evaluation, which assesses model at the process level or the outcome level.

3.1 Agentic Foundational MLLMs

Early foundational MLLMs [2, 194, 195, 196, 197, 198, 199, 200, 201] demonstrated the ability to jointly process and align images and text, achieving strong performance on a wide range of visual understanding tasks such as visual question answering [202, 203, 204], optical character recognition [205, 206], and table understanding [207, 208]. These advances mark a transformative milestone in the multimodal field, positioning MLLMs as versatile multimodal systems capable of tackling a broad spectrum of tasks.

From an architectural perspective [209, 210], foundational MLLMs can be broadly categorized into two types: **dense MLLMs**, which activate all parameters during inference, and **Mixture-of-Experts (MoE) MLLMs**, which incorporate multiple experts but activate them sparsely. With the advent of agentic MLLMs, there has been an increasing trend toward MoE architectures, as multiple experts offer better support for adaptive reasoning and dynamic tool invocation. In the following, we review recent progress in dense MLLMs and MoE MLLMs separately.

Dense MLLMs: Dense models are the classic architecture for MLLMs [2, 43, 45, 48, 198, 211], in which a single expert (i.e., a Feed-Forward Network) is employed and all parameters are activated for every input token. The forward computation is given by:

$$h^{(l+1)} = f(W^{(l)}h^{(l)} + b^{(l)}) \quad (7)$$

$$f(h) = \sigma(W_2 \sigma(W_1 h + b_1) + b_2) \quad (8)$$

where $h^{(l)}$ denotes the input at layer l , $W^{(l)}$ and $b^{(l)}$ are the corresponding weight matrices and bias terms, and $f(\cdot)$ represents the feed-forward transformation with non-linear activation $\sigma(\cdot)$. Each forward pass utilizes the full set of weights across all layers. This design is straightforward, making optimization and deployment easy and stable.

Early pioneering open-source works on dense MLLMs, such as LLaVA [197], Flamingo [195], and BLIP-2 [194], laid the foundation for multimodal understanding. More recently, a series of follow-up studies, such as Qwen2.5-VL [2], MiniCPM-V 4.5 [5], MiMoVL [46], and Key-VL-1.5 [42], have further advanced the general multimodal understanding capabilities of dense MLLMs by leveraging more powerful language models [43, 212, 213], scaling up training data [202, 203, 204, 214], and adopting improved optimization techniques [215, 216, 217, 218].

MoE MLLMs: To expand model capacity (i.e., model size) without incurring prohibitive computational costs, many foundational MLLMs adopt a Mixture-of-Experts (MoE) architecture [219, 220, 221, 222, 223]. In this design, a sparse activation mechanism ensures that only a small subset of experts is selected for each token. A trainable gating network dynamically determines the routing of inputs to experts, allowing the model to scale to billions or even trillions of parameters while keeping the per-token

computational cost comparable to that of smaller dense architectures. Such a mechanism enables specialization among experts, improves efficiency during inference, and facilitates handling of diverse multimodal tasks. Formally, the forward computation can be expressed as:

$$h^{(l+1)} = \sum_{i=1}^K g_i(x) f_i(h^{(l)}) \quad (9)$$

$$f_i(h) = \sigma(W_{2,i} \sigma(W_{1,i}h + b_{1,i}) + b_{2,i}) \quad (10)$$

$$g_i(x) = \frac{\exp(w_i^\top x)}{\sum_{j=1}^K \exp(w_j^\top x)} \quad (11)$$

where $f_i(\cdot)$ denotes the i -th Feed-Forward Network expert, $g_i(x)$ is the gating function that assigns routing weights to each expert, and $\sigma(\cdot)$ is a non-linear activation function. In practice, only the top- k experts with the highest gating weights are activated, ensuring sparse computation and improved efficiency. This makes one large model act like many specialized ones, better supporting varying levels of reasoning effort [224, 225] and diverse agentic behaviors through adaptive expert selection [213, 224].

Recent work, such as Deepseek-VL2 [56], which adopts DeepSeekMoE [226] as its language model, has demonstrated strong visual capabilities. GLM-4.5V [47] contains a total of 106B parameters, but only 12B are activated during inference, substantially enhancing its reasoning capabilities. Other studies including Kimi-VL [61], Gemini-2.5 [60], and Step-3 [59] have also leveraged MoE architectures to further enhance their performance on complex tasks. Moreover, GPT-oss [224], an MoE-based LLM, supports varying levels of reasoning effort and possesses native agentic capabilities. Building on this foundation, Intern-VL-3.5 [58] extends GPT-oss into the vision-language domain.

3.2 Agentic Action Space

Leveraging natural language as an interaction medium, MLLMs ground the definition of the action space in linguistic form, enabling the flexible and interpretable specification and execution of diverse actions. Such actions may include reasoning, reflection, memory, various tools invocation, virtual and physical environment interaction, etc. We summarize two approaches for embedding different actions into MLLMs, which are introduced in detail below.

- **Specific Tokens.** Some studies [23, 26, 101] define different actions using distinct special tokens, such as `<action_1> ... </action_1>` and `<action_2> ... </action_2>`, where the content between the action tokens specifies the corresponding operation.
- **Unified Tokens.** Other studies [24, 224] adopt a more unified approach by invoking actions with a generic `<action>...</action>` token, within which a JSON-like structure specifies the tool to be called. For example: `<action>{ 'action_name': 'action_1', 'content': '...' }</action>`.

At each state, agentic MLLMs reason over possible actions and select the best one that optimizes task completion, empowering autonomous decision-making and problem-solving capabilities that go far beyond a simple query-response chatbot.

3.3 Agentic Continual Pre-training

Agentic Continual Pre-training (Agentic CPT) [227] equips MLLMs with the ability to continually integrate new, up-to-date knowledge from diverse domains while enhancing their planning and tool-use capabilities, all without forgetting previously acquired knowledge [228, 229, 230, 231]. By reducing optimization conflicts in subsequent alignment stages, agentic CPT significantly improves overall agentic performance. The training data in this stage typically consists of large-scale synthetic corpora, and the optimization objective is based on Maximum Likelihood Estimation:

$$\mathcal{L}_{MLE}(\theta) = - \sum_{t=1}^T \log p_\theta(x_t | x_{<t}), \quad (12)$$

where x_t denotes the target token at time step t , $x_{<t}$ represents the preceding sequence of tokens, and p_θ is the conditional probability distribution over the next token parameterized by θ .

3.4 Agentic Supervised Fine-tuning

Agentic Supervised Fine-tuning (Agentic SFT) is typically introduced as an initialization stage before reinforcement learning [232, 233, 234, 235, 236, 237], providing a strong policy prior by leveraging high-quality datasets. These datasets contain detailed agentic trajectories, often synthesized through reverse engineering [233], graph-based synthesis [234, 235], and formalized task modeling [236]. The goal of agentic SFT is to align the model with action execution patterns, specifying what actions to perform and how to carry them out effectively. The optimization objective of agentic SFT remains Maximum Likelihood Estimation, consistent with Agentic CPT, though the two stages differ in both their data characteristics and training purposes.

3.5 Agentic Reinforcement Learning

Agentic Reinforcement Learning (Agentic RL) is a post-training paradigm that leverages exploration and reward-based feedback to refine agentic capabilities. Its core objective is to maximize the expected cumulative reward by iteratively refining planning processes and optimizing decision policies. We next introduce two classic RL algorithms widely used in Agentic RL, i.e., PPO [238] and GRPO [185].

Proximal Policy Optimization (PPO). PPO [238] is an actor-critic RL algorithm for aligning models with desired behaviors, which refines the policy through iterative updates that promote exploration while constraining excessive deviation from the previous policy. This balance is achieved via a clipped objective, which stabilizes optimization and mitigates the risk of performance collapse. Formally, given a policy π_θ , a previous policy $\pi_{\theta_{old}}$, and an advantage estimator A_t , the PPO objective is defined as:

$$\begin{aligned} \mathcal{J}_{PPO}(\theta) = & \mathbb{E}_{(I,T) \sim p_D, o \sim \pi_{\theta_{old}}} \\ & \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left(\frac{\pi_\theta(o_t | I, T)}{\pi_{\theta_{old}}(o_t | I, T)} A_t, \text{clip} \left(\frac{\pi_\theta(o_t | I, T)}{\pi_{\theta_{old}}(o_t | I, T)}, \right. \right. \\ & \left. \left. 1 - \epsilon, 1 + \epsilon \right) A_t \right). \end{aligned} \quad (13)$$

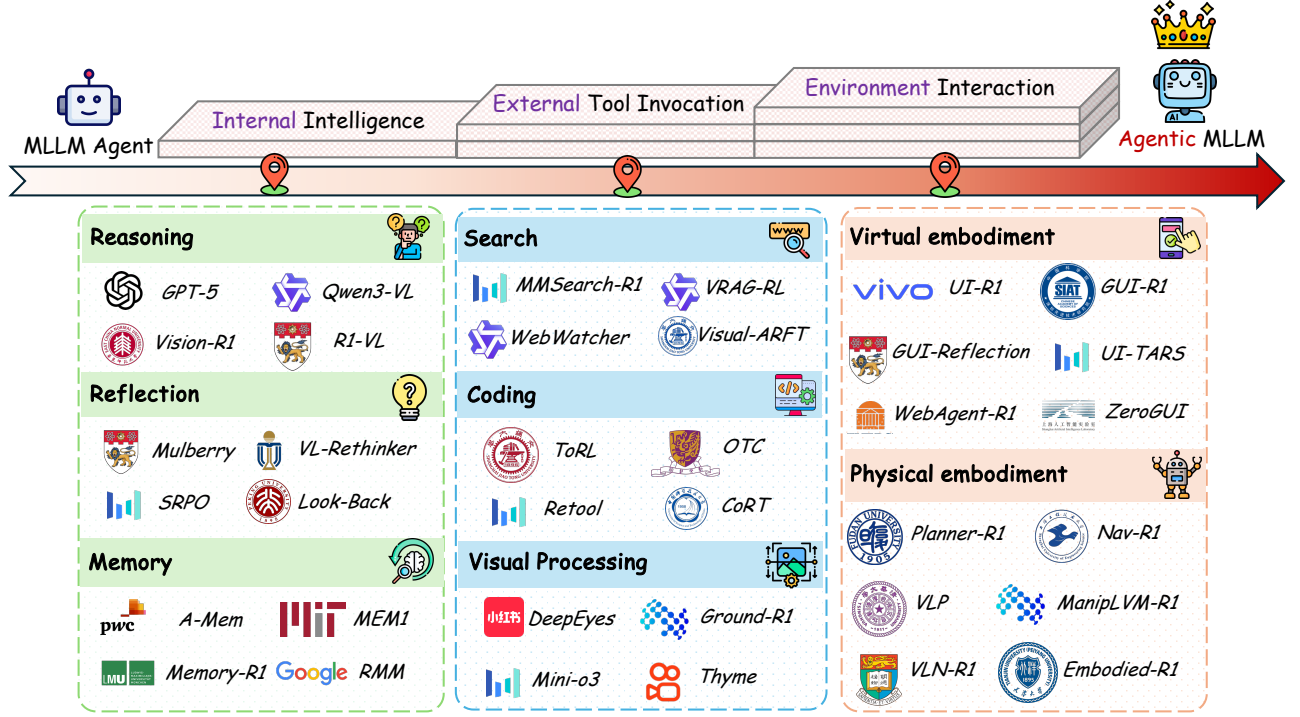


Fig. 3: The capability evolution from MLLM Agents to Agentic MLLMs: taxonomy and representative works across internal intelligence, external tool usage, and environmental interaction.

where ϵ is the clipping parameter that bounds policy updates, and A_t is the advantage, often estimated using Generalized Advantage Estimation (GAE). To further encourage linguistic coherence and mitigate reward hacking, a KL divergence penalty relative to a reference model π_{ref} is commonly added to the reward:

$$r_t = r_\varphi(q, o_{\leq t}) - \beta \log \frac{\pi_\theta(o_t | q, o_{< t})}{\pi_{\text{ref}}(o_t | q, o_{< t})}, \quad (14)$$

where r_φ denotes the reward model and β controls the regularization strength.

Group Relative Policy Optimization (GRPO). GRPO is a simplified variant of PPO that removes the need for a separate value function. It estimates the baseline directly from rollouts, reducing the cost of training a value model while maintaining stable policy updates. For each question q , GRPO samples a group of responses $\{o_1, o_2, \dots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$, with rewards $\{R_1, R_2, \dots, R_G\}$ assigned by rules or models. The rewards are then normalized by subtracting the group mean and dividing by the standard deviation to obtain the relative advantage for each response:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^G)}{\text{std}(\{R_j\}_{j=1}^G)}. \quad (15)$$

Based on these normalized advantages, the training objec-

tive is defined as:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = & \mathbb{E}_{(I, T) \sim p_D, o \sim \pi_{\theta_{\text{old}}}(\cdot | I, T)} \\ & \left[\frac{1}{n} \sum_{i=1}^n \min \left(\frac{\pi_\theta(o_i | I, T)}{\pi_{\theta_{\text{old}}}(o_i | I, T)} \hat{A}_i, \text{clip} \left(\frac{\pi_\theta(o_i | I, T)}{\pi_{\theta_{\text{old}}}(o_i | I, T)} \right), \right. \right. \\ & \left. \left. 1 - \epsilon, 1 + \epsilon \right) \hat{A}_i \right] - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \end{aligned} \quad (16)$$

where \hat{A}_i is the normalized advantage of candidate o_i , π_θ is the current policy, $\pi_{\theta_{\text{old}}}$ denotes the previous policy, π_{ref} is a reference policy for KL regularization, and ϵ and β control clipping and regularization strength, respectively.

3.6 Agentic Evaluation

Agentic MLLMs generate long-horizon action trajectories when solving complex problems. Accordingly, the evaluation can be categorized into two complementary dimensions: process evaluation and outcome evaluation.

Process Evaluation. This dimension focuses on whether the agentic MLLM can generate accurate intermediate processes, such as precise reasoning steps [239, 240, 241] or appropriate tool invocations [152, 242, 243]. It assesses the logical consistency of reasoning paths and the appropriateness of tool usage, thereby reflecting the transparency, reliability, and robustness of the intermediate process.

Outcome Evaluation. This dimension measures the ability of agentic MLLMs to produce accurate and helpful solutions across diverse downstream tasks [156, 244, 245, 246]. It reflects their generalization ability, and problem-solving competence as agentic systems.

Together, these two dimensions provide a comprehensive framework for evaluating agentic MLLMs, capturing

both the quality of their intermediate processes and the effectiveness of their final outcomes.

4 AGENTIC MLLM

In this section, we categorize agentic MLLMs into three core components: **internal intelligence** (Section 4.1), **external tool invocation** (Section 4.2), and **environmental interaction** (Section 4.3), as illustrated in Figure 3. First, internal intelligence constitutes the cognitive core of agentic MLLMs, comprising long-chain reasoning (Section 4.1.1), reflection (Section 4.1.2), and memory (Section 4.1.3). Internal intelligence enables the model to construct coherent chains of reasoning and strategic plans, orchestrating subsequent actions to accomplish tasks step by step. Second, under the coordination of internal intelligence, agentic MLLMs can proactively invoke various external tools to acquire required information (Section 4.2.1), execute code for complex computations (Section 4.2.2), and process visual representations to strengthen reasoning (Section 4.2.3). This human-like tool use substantially extends their problem-solving capabilities beyond intrinsic knowledge. Finally, with deliberate planning and tool use, agentic MLLMs interact with both virtual (Section 4.3.1) and physical environments (Section 4.3.2). Through such interactions, agentic MLLMs can perceive external environments and receive feedback, enabling dynamic adaptation in real-world deployments.

4.1 Agentic Internal Intelligence

Agentic internal intelligence denotes the capacity of a model to deliberately organize and coordinate actions in pursuit of a goal, forming the cornerstone of effective task execution. For MLLMs, achieving such internal intelligence relies on the integration of three complementary abilities: **reasoning**, **reflection**, and **memory**. These abilities collectively enable the model to coherently construct, validate, and refine its decision-making process, maintaining consistency across extended agentic trajectories. To this end, this section reviews recent approaches to advancing internal intelligence in MLLMs along these three dimensions. A summary of internal intelligence method is provided in Table 1.

4.1.1 Agentic Reasoning

Agentic reasoning in MLLMs refers to the deliberate generation of intermediate reasoning steps prior to producing a final answer, a process that substantially enhances their capacity to tackle complex problems [182, 184]. Current efforts to strengthen reasoning capabilities can be broadly categorized into three learning paradigms: **prompt-based reasoning**, **SFT-based Reasoning**, and **RL-based reasoning**. Each paradigm is introduced in the following.

Prompt-based Reasoning. Prompt-based approaches guide MLLMs to generate explicit intermediate reasoning steps by incorporating instructions such as “Let us solve the problem step by step” [255, 256]. This strategy encourages models to articulate multi-step reasoning trajectories before arriving at a final answer and has been shown to improve performance on complex tasks across diverse domains.

Building on this foundation, subsequent work has extended prompt-based CoT reasoning along both depth and

breadth. Best-of-N (BoN) methods independently generate multiple reasoning paths and then select the best one using either a reward model [257, 258, 259] or heuristic scoring functions [260, 261]. Representative studies such as VisualPRM [257], MM-PRM [258], and RM-R1 [259] train specialized reward models to better evaluate and select reasoning trajectories. Tree search methods [262, 263] further extend CoT by expanding reasoning paths into tree structures, allowing structured exploration beyond linear chains. VisuoThink [264], for instance, enables multimodal slow thinking through progressive visual-textual reasoning, while incorporating test-time scaling via look-ahead tree search. Furthermore, Monte Carlo Tree Search (MCTS) [265] introduces a principled balance between exploration and exploitation by progressively expanding promising branches through stochastic rollouts and statistical evaluation. Building on this, AStar [266] applies MCTS-derived thought cards to achieve more structured reasoning at test time.

Despite their empirical successes, prompt-based methods remain fundamentally constrained by the fixed knowledge encoded in model parameters and the limited search space available at inference. These limitations restrict their scalability and robustness when applied to more challenging, open-ended tasks.

SFT-based Reasoning. Supervised Fine-Tuning (SFT) on long-chain reasoning datasets compels MLLMs to learn reasoning abilities by minimizing the MLE loss over annotated reasoning traces. The central challenge lies in constructing high-quality reasoning datasets. We broadly categorize these approaches by their synthesis methodologies and introduce them below.

Direct distillation is a simple yet widely used method that generates reasoning paths directly from stronger teacher models, exemplified by LLaVA-Reasoner [267], MAMMO-TH-VL [268], and MAVIS [143]. Structured distillation decomposes the reasoning process into predefined modules to reduce question complexity, and then instructs the powerful model to generate each component in sequence; for example, LLaVA-CoT [144] partitions reasoning into four stages: summary, caption, reasoning, and conclusion. Tree distillation treats each reasoning step as a node in a tree, forcing the model to generate and explore multiple branches before pruning less promising ones to obtain higher-quality reasoning traces. Mulberry [19] introduces collective learning into MCTS to more effectively search reasoning and reflection trajectories.

Recent works [269, 270, 271, 272, 273, 274, 275, 276, 277] utilize these reasoning datasets to fine-tune MLLMs, advancing the development of reasoning MLLMs. Nevertheless, the reliance on high-quality CoT reasoning paths and the constrained learning mechanism of SFT, which often ties MLLMs to fixed reasoning patterns, remains a major challenge for achieving generalizable reasoning.

RL-based Reasoning. A major breakthrough in MLLM reasoning was marked by efforts such as OpenAI o1 [182] and DeepSeek R1 [184], which applied large-scale reinforcement learning and achieved transformative gains. By leveraging exploration and feedback signals, RL optimizes reasoning trajectories, allowing MLLMs to reason in a more flexible, adaptive, and dynamic manner. For long-chain reasoning, reward modeling in RL is typically divided into

TABLE 1: Summary of agentic internal intelligence, grouped into three categories: Reasoning, Reflection and Memory.

Reasoning	Fine-tuning	Reward Modeling	Contribution
Vision-R1 [15] [code]	SFT+RL	Outcome + Rule	Introduce progressive thinking suppression training within GRPO to progressively optimize the model.
MM-Eureka [67] [code]	SFT+RL	Outcome + Rule	Introduce high-quality MM-K12 with online filtering and a two-stage training strategy to improve stability.
Skywork R1V2 [76] [code]	SFT+RL	Outcome + Rule	Propose selective sample buffer to tackle GRPO’s vanishing advantages by prioritizing high-value samples.
Video-R1 [73] [code]	SFT+RL	Outcome + Rule	Construct 165K cold start and 260K RL dataset; Propose T-GRPO to explicitly encourage temporal reasoning in videos.
LongVILA-R1 [89] [code]	SFT+RL	Outcome + Rule	Introduce 104K long-video QA pairs with reasoning annotations and a two-stage pipeline of CoT-SFT and RL.
ThinkLiteVL [65] [code]	RL	Outcome + Rule	Repurpose MCTS to identify challenging yet solvable examples that enhance RL effectiveness in low-data regimes.
R1-ShareVL [66] [code]	RL	Outcome + Rule	Expand the question space and shares reasoning trajectories and rewards across variants to mitigate sparse rewards.
EchoInk-R1 [247]	RL	Outcome + Rule	A GRPO framework for audio-image QA, showing reflection by revisiting and refining responses under ambiguity.
Infri-MMR [248] [code]	RL	Outcome + Rule	A curriculum learning activating reasoning with text, adapting with captions and enhancing with caption-free data.
NoisyRollout [249] [code]	RL	Outcome + Rule	Augment RL by mixing clean and distorted trajectories with noise annealing to improve exploration and robustness.
VL-Cogito [74] [code]	RL	Outcome + Rule	Curriculum RL with difficulty soft weighting and dynamic length rewards to balance efficiency and correctness.
WeThink [69] [code]	RL	Outcome + Model	Present a hybrid reward combining rule-based verification and model-based assessment to optimize RL across tasks.
R1-VL [16] [code]	SFT+RL	Process + Rule	Introduce two rewards to help models cover key intermediate steps while maintaining structural and logical consistency.
SophiaVL-R1 [75] [code]	SFT+RL	Process + Model	Introduce process-level rewards by a trained reward model and using Trust-GRPO to weight their reliability.
Perception-R1 [250] [code]	RL	Process + Model	Propose a visual perception reward, judged by an LLM for annotation-response consistency.
GRPO-CARE [78] [code]	RL	Process + Model	Consistency-aware learning with correctness rewards and an adaptive consistency bonus for coherent reasoning.
Reflection	Trigger Type	Reflection Granularity	Contribution
VLAAR-Thinker [33] [code]	Implicit	Step level	Demonstrate that GRPO training induces reflection, evidenced by the frequency of four “aha” expressions.
MM-Eureka [67] [code]	Implicit	Step level	RL optimization induces reflection in MLLMs without explicit incentives.
FRANK [85] [code]	Implicit	Step level	Propose hierarchical weight merging of a MLLM and a reasoning-specialized LLM, revealing emergent reflection.
Mulberry [19] [code]	Explicit	Step level	Leverage CoMCTS to build reflective reasoning paths by incorporating negative sibling nodes into trajectories.
Vision-R1 [15] [code]	Explicit	Step level	Introduce a cold-start dataset, vision-r1-cold, featuring a higher frequency of reflective markers.
VL-Rethinker [81] [code]	Explicit	Step level	Explicit “rethinking triggers” during rollouts, guiding VLMs toward strategic reflection.
Gthinker [80] [code]	Explicit	Step level	Propose a reasoning pattern that grounds in visual cues and iteratively reinterprets them to resolve inconsistencies.
R ³ V [79] [code]	Explicit	Response level	Iteratively generate positive/negative solutions, apply self-reflection to refine flaws, and select superior reasoning paths.
SRPO [20] [code]	Explicit	Response level	Proposes a two-stage framework to enhance reasoning with reflection-focused data and a reflection-aware GRPO reward.
Look-Back [82] [code]	Explicit	Response level	Introduce an implicit method enabling MLLMs to self-reflect by re-focusing on visual inputs during reasoning.
LongVIL [84] [code]	Explicit	Response Level	An agent with plan and code reflection to refine actions and code, ensuring temporal-spatial coherence and correctness.
Memory	Memory Type	Mechanism	Contribution
BLIP-2 [194] [code]	Contextual	Token Compression	Leverage a two-stage pretrained Querying Transformer to bridge the modality gap and compress visual tokens.
Dense Connector [251] [code]	Contextual	Token Compression	Use a parameter-free connector layer to compress visual tokens, accelerating inference while preserving performance.
Qwen2.5-VL [2] [code]	Contextual	Token Compression	An MLP compresses adjacent visual patch features into the text embedding space for efficient vision-language fusion.
LongRoPE [88] [code]	Contextual	Window Extension	Extend window to 2048K by non-uniform interpolation search, progressive extension training, LongRoPE readjustment.
LongLM [252] [code]	Contextual	Window Extension	Extend context window by constructing bi-level attention information: the grouped attention and the neighbor attention.
LongVA [253] [code]	Contextual	Window Extension	Extrapolate LLM’s context length, enabling MLLMs to comprehend orders of magnitude more visual tokens.
LongVILA [89] [code]	Contextual	Window Extension	Upgrade VLMs to support long context understanding by long context extension and long video SFT.
S ² CAN [254]	External	Heuristic-driven	A memory-augmented framework enhancing surgical context understanding with direct and indirect memories.
MA-LMM [86] [code]	External	Heuristic-driven	Design distinct visual and query memory banks to separately manage information from different modalities.
MovieChat [87] [code]	External	Heuristic-driven	Combine a sliding-window short-term memory with a compact long-term memory to consolidate video tokens.
MemoryBank [90]	External	Heuristic-driven	Evolve memories, adapt to users, and use an Ebbinghaus-inspired mechanism to forget or reinforce information.
MemTool [91]	External	Heuristic-driven	Short-term memory for tool/context control in multi-turn conversations with autonomous, workflow and hybrid modes.
A-Mem [21] [code]	External	Reasoning-driven	A Zettelkasten-inspired memory system building evolving knowledge networks via dynamic indexing and linking.
MEM1 [93] [code]	External	Reasoning-driven	An RL framework maintaining constant memory in multi-turn tasks via compact updates and redundancy reduction.
Memory-R1 [22]	External	Reasoning-driven	An RL-based memory manager and answer agent for adaptive external memory management beyond static heuristics.
RMM [94]	External	Reasoning-driven	Long-term dialogue with Prospective Reflection for memory and Retrospective Reflection for RL-based refinement.
M+ [95] [code]	External	Reasoning-driven	A memory-augmented model with long-term memory and a co-trained retriever for dynamic retrieval during generation.

two paradigms: outcome-based rewards, which evaluate only the final answers, and process-based rewards, which additionally assess intermediate reasoning steps. Both can be assigned rewards through either rule-based heuristics or specialized reward models. In the following, we review representative methods according to their reward formulations.

- **Outcome reward modeling** assigns rewards based solely on final answer correctness, ignoring the intermediate reasoning process. It is simple to implement and has attracted widespread attention, particularly following the success of DeepSeek-R1 [184], which employed rule-based reward computation to mitigate reward-model hacking [278, 279] and to lower the need for additional training resources. Subsequent work [68, 70, 280, 281, 282, 283] extended outcome-based RL to the multimodal domain. Early work Vision-R1 [15] introduces progressive thinking suppression training into GRPO, mitigating token explosion and improving training stability. Recent works enhance multimodal reasoning capabilities through techniques such as high-quality data selection [65, 74] and data augmentation [66, 249], addressing advantage vanishing [66, 76, 81, 284] and curriculum learning [74, 248, 285]. Beyond rule-based judgment, WeThink [69] has also introduced reward models to verify the correctness of final answers.
- **Process reward modeling** extends outcome-based rewards by incorporating supervision at the intermediate step level in addition to outcome-level evaluation, guiding intermediate reasoning steps to improve

the quality and robustness of the reasoning process. R1-VL [16] introduces rule-based process rewards by matching extracted keywords from reasoning steps to predefined rules, enabling finer-grained control and alleviating advantage vanishing. Other works, such as Perception-R1 [250], SophiaVL-R1 [75], and GRPO-CARE [78], employ specialized process reward models to score intermediate steps, improving reasoning reliability, coherence, and consistency.

In summary, outcome-based rewards are simple to implement and efficient to scale, but they overlook the reasoning process. In contrast, process-based rewards provide finer-grained supervision that improves reliability and coherence, though they demand more complex design, incur higher computational costs, and remain vulnerable to reward model hacking.

4.1.2 Agentic Reflection

MLLMs are inherently constrained by the autoregressive paradigm, where errors are irreversible and tend to accumulate over time. Drawing inspiration from its central role in human cognition, reflection has been introduced into LLMs as a mechanism to overcome this limitation. Recent studies [19, 79, 286] demonstrate that reflective strategies enable models to verify and refine their responses, thereby enhancing robustness, mitigating hallucinations, and supporting more effective agentic internal intelligence. The approaches for inducing reflection can be categorized into explicit and implicit methods.

Implicitly Induced Reflection. Studies such as DeepSeek-R1 [184] have observed that models can exhibit emergent reflective behaviors after reinforcement learning. These reflective behaviors are not explicitly induced, but rather emerged organically through interaction with the reinforcement learning exploration. Similar emergent reflections have also been reported in MLLMs, as shown by MM-Eureka [67] and VLAA-Thinker [33].

Explicitly Induced Reflection. Subsequent research [19, 20, 80] introduces mechanisms that explicitly induce reflective behaviors in MLLMs. These methods can be broadly divided into two categories: **response-level** reflection, which is applied after the model generates a complete response, and **step-level** reflection, which is introduced during intermediate reasoning steps.

- **Response-level reflection.** In this setting, reflection is triggered only after the model generates a complete response, which can be formalized as:

$$response = r^- + \rho + r^+, \quad (17)$$

where r^- denotes the initial flawed response, r^+ represents the refined response, and ρ is the reflection prompt linking the two. Representative methods include R³V [79], which fosters reflective capability by iteratively generating positive and negative solutions, applying self-reflection losses to refine flawed rationales, and selecting superior reasoning paths. SRPO [20] introduced a two-stage RL framework that leverages reflection-enhanced data and reflection-aware GRPO rewards to incentivize reflective behaviors.

- **Step-level reflection.** In this setting, reflection is interleaved between intermediate reasoning steps so that each draft step is critiqued and revised before proceeding, formalized as:

$$response = s_1 + s_2^- + \rho + s_2^+ + \dots + s_n, \quad (18)$$

where s_t^- denotes the t -th initial flawed reasoning step, s_t^+ represents the revised step after reflection, and ρ indicates the reflection prompt inserted between consecutive steps. Mulberry [19] exemplifies this paradigm by employing collective MCTS to construct reflective reasoning paths, explicitly incorporating negative sibling nodes to incentivize reflection. VL-Rethinker [81] advances this direction by designing explicit rethinking triggers during rollouts, guiding MLLMs toward more strategic reflection.

4.1.3 Agentic Memory

Memory plays a pivotal role in advancing MLLMs beyond the limitations of the fixed and limited context window. By retaining and leveraging past information, it enables models to maintain continuity across sessions, and support more coherent internal intelligence over long-horizon interactions. In this section, we divide memory into **contextual** and **external** memory systems for detailed discussion. The summary of agentic memory research is provided in Figure 1

Contextual Memory. Contextual memory refers to directly concatenating past information into the current context window, providing a simple yet effective way to leverage history for response generation. However, the fixed

context length imposes strict limits, motivating two primary strategies: **token compression** and **window extension**.

- **Token compression.** This strategy reduces the number of tokens by condensing input representations, thereby indirectly increasing the effective capacity of the context window. Parametric methods typically employ a Query Transformer (Q-Former) to downsample high-dimensional features into a smaller set of informative learnable tokens. Representative works include Flamingo [195], BLIP-2 [194], and Video-LLaMA [287], which use Q-Former as a vision-language bridge to efficiently compress visual inputs. In contrast, non-parametric approaches rely on traditional pooling operations (e.g., average pooling or max pooling) [288]. Such methods have been explored in PLLaVA [289] and Dense Connector [251], where pooling is applied to compress multimodal inputs without introducing additional learnable parameters.
- **Window Extension.** Unlike token compression, which enlarges context capacity indirectly, another line of work focuses on directly extending the context window. LongRoPE [88] expands the original context length from 128k to 2048k tokens through a progressive extension strategy. In the multimodal domain, LongVILA [89] and LongVA [253] extend the context window to handle inputs exceeding 2,000 video frames, supporting long-horizon temporal reasoning.

External Memory Systems. Some studies [90, 290] extend memory beyond the internal context window by incorporating external modules for storing and retrieving information. Based on their mechanisms, these approaches can be broadly divided into heuristic-driven and reasoning-driven memory systems.

- **Heuristic-driven memory.** Early external memory systems relied on static, rule-based pipelines with predefined strategies for storing, updating, and retrieving information. For example, MemoryBank [90] and MemGPT [290] use specialized prompts to manage textual memory, while MovieChat [87] and MovieChat+ [291] introduce both short-term and long-term modules to process videos exceeding 10K frames. Similarly, MA-LMM [86] maintains separate memory banks for visual and query information. Although effective in constrained domains, these systems depend on fixed heuristics, which limit adaptability in dynamic and open-ended environments.
- **Reasoning-driven memory.** Building on these foundations, recent research has advanced toward reasoning-driven memory systems that autonomously store, update, and utilize memory in a more dynamic and task-driven manner. A-Mem [21] introduces an agentic memory framework inspired by the Zettelkasten method, allowing LLM agents to dynamically organize and evolve interconnected memory nodes for more adaptive, context-aware reasoning. Mem0 [92] proposes a scalable memory-centric architecture that dynamically manages salient information, with a graph-based variant to capture relational structures, yielding superior long-term conversational coherence. MemTool [91] focuses on short-term memory, enabling agents to dy-

TABLE 2: Summary of agentic external tool invocation, grouped into three categories: Search, Coding and Visual Processing.

Agentic Search	Fine-tuning	Search Modalities	Contribution
MMSearch [192] [code]	Prompt-based	T, I	Introduce multimodal AI search engine pipeline that equips MLLMs with multimodal search capabilities.
Search-R1 [101] [code]	RL	T	RL framework to autonomously generate multi-turn search queries stabilized by retrieved token masking and outcome reward.
VRAG-RL [96] [code]	SFT+RL	I	Visual action space with cropping–scaling for info gathering and a reward uniting query rewriting and retrieval performance.
Visual-ARFT [26] [code]	RL	T, I	Enable MLLMs to flexibly reason by browsing websites for real-time information and coding adaptive image manipulations.
MM-Search-R1 [23] [code]	SFT+RL	T, I	Learn when and how to perform image–text search by SFT and RL, guided by outcome-based rewards with a search penalty.
M2IO-R1 [99]	SFT+RL	T, I	A MRAG framework supporting multimodal I/O with controllable, semantically aligned image selection and placement.
Patho-AgenticRAG [100] [code]	SFT+RL	T, I	Page-level embedding database for text–image retrieval with reasoning, decomposition, and multi-turn search in diagnostics.
WebWatcher [24] [code]	SFT+RL	T, I	Leverage synthetic multimodal trajectories for efficient cold-start, enabling tool use and improved generalization via RL.
Agentic Coding	Fine-tuning	Application	Contribution
Posterior-GRPO [292]	RL	Programming	Introduce Posterior-GRPO and tailored reward models to guide intermediate reasoning for more accurate code generation.
R1-Code-Interpreter [293] [code]	SFT+RL	Programming	Achieve successful multi-turn interleaved textual reasoning and code generation across multiple tasks.
ToRA [103] [code]	SFT	Mathematics	A pioneering line of work that integrates external coding tools into the textual reasoning process.
MathCoder [104] [code]	SFT	Mathematics	Present MathCodeInstruct, a high-quality SFT dataset, and MathCoder, a family of models for mathematical reasoning.
rStar-Math [105] [code]	SFT	Mathematics	Propose a self-evolution framework that integrates an MCTS-based data synthesis method with a process preference model.
ToRL [106] [code]	RL	Mathematics	Achieve tool-integrated reasoning on challenging mathematical problems through reinforcement learning.
Retool [25] [code]	SFT+RL	Mathematics	Constructs an outcome-driven RL framework for multi-turn tool invocation and long-form reasoning.
OTC [294]	RL	Mathematics	Incentivize models to solve tasks correctly using minimal tool interactions via Optimal Tool Call-controlled Policy Optimization.
CoRT [107] [code]	SFT+RL	Mathematics	Propose a hint-engineering strategy that employs targeted prompts to guide reasoning and suppress redundant text generation.
rStar2-Agent [295] [code]	SFT+RL	Mathematics	Introduce an efficient RL infrastructure with a tailored GRPO-RoC strategy, enabling a powerful agentic reasoning model.
MedAgentGym [296] [code]	SFT+RL	Healthcare	Advance coding-based medical reasoning by constructing a training environment that spans diverse biomedical scenarios.
ML-Agent [297] [code]	SFT+RL	Machine Learning	Pioneers agentic machine learning engineering through online reinforcement learning in interactive environments.
Agentic Visual Processing	Fine-tuning	Processing Type	Contribution
DeepEyes [27] [code]	RL	Cropping	Introduce a tool-use-oriented data selection mechanism and reward strategy to foster “thinking with images” capabilities.
Ground-R1 [108]	RL	Cropping	Propose a reinforcement learning framework that Present scalable grounded visual reasoning without costly annotations.
Active-O3 [109] [code]	RL	Cropping	Propose an RL framework that equips MLLMs with efficient active perception capabilities for tasks like small-object grounding.
Chain-of-Focus [110] [code]	SFT+RL	Cropping	Enable MLLMs to perform adaptive region focusing and zooming through a two-stage training pipeline.
Pixel-Reasoner [111] [code]	SFT+RL	Cropping	Propose pixel-space reasoning, a novel framework that equips MLLMs with visual operations (e.g., zoom-in, select-frame).
VLM-R3 [112]	SFT+RL	Cropping	Equip MLLMs with region recognition and reasoning capabilities via Region-Conditioned Reinforcement Policy Optimization.
Mini-o3 [28] [code]	SFT+RL	Cropping	Enable deep multi-turn reasoning with tool interactions, and achieves leading performance on complex visual search tasks.
OpenThinkIMG [113] [code]	SFT+RL	Manipulation	Build a tool-augmented agentic MLLM with adaptive tool-use capabilities for complex chart reasoning tasks.
Thyme [114] [code]	SFT+RL	Manipulation	Present MLLMs to autonomously generate and execute image processing and computational code.
VILASR [298] [code]	SFT+RL	Manipulation	Equip MLLMs with elementary drawing operations (e.g., bounding boxes, auxiliary lines) to enhance spatial reasoning.
REVPIT [115] [code]	SFT+RL	Manipulation	Enhance MLLMs’ visual perception and reasoning by training them to dynamically leverage a suite of specialized visual tools.
VPRL [116] [code]	RL	Generation	Propose visual planning that replaces text-based reasoning with coherent image sequences generated by a large vision model.

namically manage tools or MCP server contexts across multi-turn conversations; it provides Autonomous, Workflow, and Hybrid modes with distinct trade-offs between efficiency and accuracy. More recently, Memory-R1 [22] introduces an RL-based framework for adaptive external memory management. It employs a Memory Manager that learns structured operations (e.g., add, update, delete, noop) and an Answer Agent that retrieves and reasons over relevant entries, enabling continuous and flexible memory beyond static, rule-based approaches.

Despite these advances, most work on agentic memory remains text-centric, leaving a notable gap in multimodal agentic memory management for future research.

4.2 Agentic External Tool Invocation

“A good tool improves the way you work. A great tool improves the way you think.”
— Jeff Duntemann

While internal intelligence equips agentic MLLMs with the ability to reason, reflect and memory, their capabilities remain intrinsically limited to the knowledge encoded in the model parameters. A natural strategy to overcome this limitation is to augment MLLMs with the ability to use external tools for problem solving. Early approaches [192, 299, 300] relied on prompt engineering to passively trigger tool use, but such methods lack the flexibility and adaptability required for novel tasks. Recent advances in agentic MLLMs have shifted this paradigm by integrating tool invocation into the reasoning process, enabling models to incorporate external tools into step-by-step reasoning and to autonomously determine when, and which tools to employ. To this end, in this section, we review how agentic MLLMs learn to reason with external tools, categorized by different tool types, including information searching, code execution, and visual processing. A summary of agentic external tool invocation method is presented in Table 2.

4.2.1 Agentic Search for Information Retrieval

In today’s rapidly evolving information landscape, a pressing requirement for intelligent systems is the ability to stay current with emerging knowledge. However, once training is complete, the knowledge space of MLLMs becomes fixed and they cannot directly handle newly emerging events or rapidly changing domains. For example, GPT-5 [301], released in August 2025, only retains knowledge up to June 2024, leaving it unable to address subsequent developments. To overcome this limitation, researchers have proposed augmenting MLLMs with web search integration [302, 303, 304] or Retrieval-Augmented Generation (RAG) [305, 306], enabling access to external knowledge sources such as the Internet or specialized databases. This integration extends their capabilities beyond static parametric knowledge and enhances adaptability to dynamic real-world contexts.

Search Agent. Traditional MLLM search agents [192, 299] often pre-define a sequential pipeline to execute search instructions and retrieve external knowledge for problem solving. For example, when presented with an up-to-date question, the agent system first reformulates the query and submits it to a search engine, then reranks the retrieved results, and finally prompts the MLLM to synthesize the information into a coherent answer for the user. Representative agent systems such as MMSearch [192] and MindSearch [307] use this paradigm, proposing structured pipeline designs to enable access to external knowledge.

Agentic Search. Agentic search leverages end-to-end reinforcement learning to equip MLLMs with the autonomy to decide both when to search and what to search for. By embedding search directly into the reasoning process, this paradigm reduces redundant queries and enables retrieval that aligns more coherently with multi-turn interactions. Training an agentic search model typically begins with curating up-to-date or knowledge-intensive questions that require external resources to answer, and constructing

corresponding question–answer pairs. Recent studies have proposed diverse strategies for building such datasets [23, 24, 26, 227, 233, 236], including reverse engineering, graph-based synthesis, and formalized task modeling. Based on these data, reinforcement learning with tailored reward functions [23, 24, 26, 101] is then employed to incentivize the model’s ability to conduct adaptive and contextually appropriate search.

Pioneering works focus on searching textual corpora. For instance, Search-R1 [23] integrates search into LLM reasoning with token-masked retrieval, interleaved multi-turn reasoning, and outcome-based rewards to stabilize RL training and enhance complex task solving. Search-o1 [102] introduces a framework that integrates agentic search into o1-like reasoning, enabling LLMs to retrieve and refine external knowledge on demand while preserving logical flow. Moreover, Visual-ARFT [26] augments multimodal understanding by integrating text search.

Building on this foundation, subsequent studies extend agentic search to multimodal information retrieval. This is achieved via custom multimodal search frameworks [96] or specialized engines such as Google SerpApi¹. VRAG-RL [96] defines a visual action space with cropping and scaling for coarse-to-fine information gathering, reinforced by a reward combining query rewriting and retrieval accuracy. MMSearch-R1 [23] integrates both image and text search tools, leveraging cold-start training and RL to teach models when and how to invoke each tool, guided by outcome-based rewards with search penalties. WebWatcher [24] further advances this line by systematically constructing search-dependent data and introducing unified special tokens to coordinate image and text search engines.

4.2.2 Agentic Coding for Complex Computations

While MLLMs have demonstrated remarkable capabilities in cross-modal vision-language tasks, they remain inherently limited in tasks requiring rigorous program synthesis, precise mathematical computation, and structured symbolic reasoning. A key development in overcoming these challenges is the emergence of agentic MLLMs, which autonomously plan, generate, and refine code-based actions through iterative program reasoning and dynamic tool utilization. Guided by the primary application domains of agentic coding, this section surveys these synergistic advances by categorizing recent work into three key areas: **program engineering**, **mathematical reasoning**, and **other domain-specific applications**.

Program Engineering. Recent research has extensively explored methods for taming LLMs to function as capable coding assistants [309, 310, 311]. The integration of RL has further augmented these capabilities, allowing for self-improvement in both code generation and execution accuracy [312, 313, 314]. One line of work utilizes outcome-based rewards, such as code execution and test case results, as direct training signals [314, 315]. In contrast, another strand of research introduces denser, process-oriented rewards that provide stepwise feedback on aspects such as code snippets and intermediate reasoning, thereby offering finer-grained

guidance during training [292, 316, 317]. Subsequent studies have expanded these efforts, broadening the scope of agentic coding to include iterative code refinement through multi-turn interactions [293], co-evolution of code generators and unit testers to improve robustness [318], and the application in advanced software engineering tasks [319, 320].

Mathematical Reasoning. Numerous studies have integrated external tools, such as computational libraries and symbolic solvers, directly into the reasoning process, a methodology now commonly termed tool-integrated reasoning. This enables models to dynamically execute code and obtain reliable numerical and symbolic solutions, significantly improving performance in complex reasoning tasks like mathematical problem-solving. Specifically, early efforts focus on building high-quality reasoning trajectories that interleave natural language reasoning with code execution, thereby stimulating the model’s capacity to autonomously generate and execute code [103, 104, 105]. As a pioneering effort, ToRA [103] first prompts advanced LLMs like GPT-4 to synthesize high-quality reasoning trajectories with tool calls for imitation learning. Subsequently, an output space shaping strategy is employed to augment the dataset with the initial model’s self-generated correct trajectories and its errors after teacher model correction. A final SFT phase on this enriched data further enhances the model’s capabilities in leveraging external tools and generating code to solve complex mathematical problems.

Fueled by recent advances in large reasoning models, leveraging RL to autonomously integrate code generation into text-centric long CoT reasoning is an emerging research trend [25, 106]. For instance, ToRL [106] employs a pure RL strategy to promote code-integrated reasoning, while ReTool [25] further enhances long-form capabilities through an outcome-driven RL framework that supports multi-turn code execution. Subsequent research has placed greater emphasis on balancing accuracy and efficiency in models that actively employ code generation for reasoning. In this vein, OTC [294] introduces an Optimal Tool Call-controlled Policy Optimization that incentivizes models to solve tasks correctly using minimal tool interactions. CoRT [107] pinpoints two primary sources of inefficiency: first, a delay in code computation caused by a default to textual CoT reasoning prior to code generation; and second, a distrust in code results, which triggers superfluous manual verification of the execution outputs. To address these challenges, CoRT introduces a hint-engineering strategy that inserts strategic prompts to steer the reasoning trajectory, thereby avoiding the overhead of futile textual reasoning.

Other domain-specific applications. Beyond the above-mentioned advancements, recent research has successfully extended agentic coding techniques to a variety of other domains, e.g., healthcare [296] and machine learning [297]. These cross-disciplinary efforts demonstrate the remarkable adaptability and impact of agentic coding, highlighting its potential to transform complex decision-making processes and operational workflows across diverse sectors.

4.2.3 Agentic Visual Processing for Thinking with Image

Recent advances demonstrate a paradigm shift in large reasoning models from text-centric approaches towards integrated multimodal reasoning, which jointly interleaves tex-

¹SerpApi: <https://serpapi.com/>

TABLE 3: Summary of agentic environment interaction, grouped into two categories: virtual and physical.

Agentic Virtual Embodiment	Fine-tuning	Learning Type	Contribution
AGUVIS [117] [code]	SFT	Offline	Integrate structured reasoning and operates autonomously as a unified vision-based GUI agent.
InfGUIAgent [118] [code]	SFT	Offline	Cultivate native hierarchical and expectation-reflection reasoning skills to enhance multi-step GUI automation.
TongUI [119] [code]	SFT	Offline	Mitigate data scarcity for GUI agents by automatically generating the GUI-Net-1M dataset from multimodal web tutorials.
UI-R1 [173] [code]	RL	Offline	Enable MLLMs to achieve significant accuracy improvements in GUI action prediction with exceptional data efficiency.
GUI-R1 [30] [code]	RL	Offline	Leverage unified action space modeling and policy optimization to dramatically enhance the generalization and data efficiency.
InfGUI-R1 [124] [code]	RL	Offline	Introduce the Actor2Reasoner framework that transforms reactive GUI agents into deliberative reasoners.
ComfyUI-R1 [120] [code]	SFT+RL	Offline	Propose a specialized reasoning model that achieves automated workflow generation through a two-stage training framework.
GUI-Reflection [121] [code]	Pretraining+SFT+RL	Online	Develop self-correction capabilities in GUI agents through automated reflection data generation and iterative online tuning.
ZeroGUI [123] [code]	RL	Online	Eliminate human annotation costs by automating task generation and reward estimation through MLLMs.
WebAgent-R1 [122] [code]	RL	Online	Achieve strong gains in multi-turn web interactions via asynchronous trajectory generation and binary reward optimization.
UI-TARS [125] [code]	Pretraining+SFT+RL	Online	Integrate innovations in screenshot perception, unified action modeling, deliberate reasoning and iterative self-improvement.
UI-TARS-2 [308] [code]	Pretraining+SFT+RL	Online	A systematic framework addressing data scalability, multi-turn RL stability, hybrid environment and unified sandbox platform.
Agentic Physical Embodiment	Fine-tuning	Task Type	Contribution
ALP [126] [code]	RL	Perception	Combine action-aware representation learning with active environmental exploration to learn robust visual representations.
EAR [127]	RL	Perception	Model visual exploration as sequential evidence gathering with an uncertainty-aware reward for open-world environments.
Wu et al. [128] [code]	SFT+RL	Planning	Incorporate R1-style reasoning to advance embodied planning performance and generalization in interactive environments.
Embodied Planner-R1 [129] [code]	RL	Planning	Introduce an RL framework with sparse completion rewards and interactive policy optimization for embodied planning.
OctoNav [130] [code]	SFT+RL	Navigation	Construct a large-scale benchmark and a unified framework with think-before-action capability for generalist navigation agents.
VLN-R1 [131] [code]	SFT+RL	Navigation	Propose a GRPO-based RL method with time-decayed rewards for continuous embodied navigation.
Nav-R1 [132] [code]	SFT+RL	Navigation	Decouple high-level planning from low-latency control and enables coherent yet highly responsive navigation.
VLP [133] [code]	RL	Manipulation	Advance a new approach to embodied manipulation via a language-conditioned preference feedback framework.
ManipLVM-R1 [134]	RL	Manipulation	Develop an RL framework with two tailored reward functions for spatial perception and trajectory matching.
Embodied-R1 [135] [code]	RL	Manipulation	Bridge the robotics perception-action gap with a pointing-centric representation and an RL-based training strategy.

tual and visual information. This evolution is often driven by the agentic invocation of tools or functions, enabling a form of “thinking with images” [321]. Based on their distinct approaches to image processing, we can roughly categorize the evolution into three main phases: thinking with **cropped** images, thinking with **manipulated** images, and thinking with **generated** images.

Thinking with cropped images: As the early open-source initiative of its kind, DeepEyes [27] effectively integrates visual information into textual chain-of-thought reasoning by leveraging the model’s inherent grounding capabilities, augmented with cropping and zoom-in functions. The training framework relies exclusively on reinforcement learning (i.e., GRPO) with tailored reward functions, eliminating the need for cold-start SFT. Concurrent works such as Ground-R1 [108] and Active-O3 [109] implement similar RL-driven concepts, differing only marginally in their use of training data and reward design. Another line of research, exemplified by methods such as Chain-of-Focus [110], Pixel-Reasoner [111], and VLM-R³ [112], employs cold-start SFT to equip models with multimodal reasoning strategies and structured output formats in advance, thereby alleviating the burden on subsequent reinforcement learning. These approaches sample their initial training data from existing datasets such as VisCoT [322], or leverage GPT-4o to curate examples based on image collections like SA-1B [323]. Remarkably, the latest Mini-o3 [28] achieves deep multi-turn exploration with tool interactions through specialized dataset construction, diverse trajectory collection, and innovative over-turn masking strategies, leading to state-of-the-art performance on challenging visual search tasks.

Thinking with manipulated images: Beyond fundamental operations such as cropping and zooming, more advanced approaches endow models with enhanced capabilities for active image manipulation. OpenThinkIMG [113] builds a tool-augmented, agentic MLLM with adaptive tool-use capabilities for complex chart reasoning tasks. The toolset encompasses both basic operations (e.g., crop, zoom-in, and draw) and powerful external models including SAM [323] and GroundingDino [324]. Thyme [114] utilizes agentic code generation to perform autonomous image editing (e.g., cropping, rotation, contrast enhancement) and mathematical computations, within its reasoning process.

This method uses a two-stage SFT and RL training paradigm and introduces GRPO with Adaptive Temperature Sampling (GRPO-ATS), which decouples text and code sampling temperatures to ensure high-fidelity code generation. VILASR [298] extends this concept to spatial intelligence, enabling the model to edit images or video frames by drawing additional bounding boxes or auxiliary lines. Experiments across multiple benchmarks confirm that this method consistently boosts spatial reasoning performance. Furthermore, ReVPT [115] incorporates a comprehensive visual toolkit, including depth estimation, zoom in, object detection, and edge detection. Empowered by cold-start SFT and RL training, it demonstrates significantly enhanced visual perception, setting a new state-of-the-art on spatial reasoning and image understanding benchmarks.

Thinking with generated images: Recently, a growing number of efforts extend reinforcement learning to image generation, leveraging it to unlock MLLM reasoning for creating high-fidelity images that are better aligned with human instructions [325, 326, 327]. In parallel, another line of research explores active image generation for enhanced visual understanding. For instance, the VPRL [116] method employs reinforcement learning to endow large vision models with visual chain-of-thought reasoning capabilities. By generating a sequence of images that provides coherent visual cues, these models achieve significant performance gains in visual planning tasks.

4.3 Agentic Environment Interaction

Beyond reasoning and tool utilization, agentic environment interaction represents the stage where MLLMs transcend static query-response paradigms and begin engaging with their surroundings. Through continuous virtual or physical interaction (i.e., executing actions, perceiving environmental changes, and integrating feedback), agentic MLLMs dynamically adjust their strategies in response to evolving contexts, enabling them to pursue long-term goals, adapt in real time, and align their behaviors with the surrounding environment. A summary of agentic environment interaction method is presented in Table 3.

4.3.1 Agentic Virtual Environment Interaction

Recent years have witnessed significant advances in agentic MLLMs capable of performing complex tasks through

graphical user interfaces (GUIs). These GUI agents, which enable autonomous interaction with digital environments, have evolved into increasingly sophisticated systems that leverage learning-based approaches to generalize across diverse applications and platforms. In this section, we categorize these systems based on their learning mechanisms: one that **learn from pre-collected GUI demonstration trajectories**, and another that **learn directly through interaction within dynamic GUI environments**. We systematically examine both categories, highlighting their representative methods, key strengths, and inherent limitations.

Learning from offline demonstration trajectories: AGU-VIS [117] introduces a large-scale GUI trajectory dataset and a two-stage training framework that decouples visual grounding from high-level planning, establishing state-of-the-art performance across offline and online GUI benchmarks. InfiGUIAgent [118] also adopts a two-stage SFT workflow that first instills core GUI grounding skills, then enhances reasoning and reflection capabilities using synthesized data. TongUI [119] addresses the critical bottleneck of limited training data for generalized GUI agents by automatically constructing a large-scale, multimodal dataset, termed GUI-Net-1M, from crawled web tutorials. By fine-tuning the Qwen2.5-VL [2] models on this dataset, the resulting TongUI agent demonstrates a substantial performance gain on standard grounding and navigation benchmarks, validating the framework’s effectiveness and the utility of the newly created resource.

Despite technical progress, the conventional SFT training paradigm exhibits a strong dependency on massive, curated datasets and hinders model generalization in unseen environments. To address this limitation, significant research efforts are devoted to integrating RL into GUI-based tasks. UI-R1 [173] first proposes an RL framework that significantly enhances GUI action prediction through policy optimization with novel rule-based action-level rewards. This approach demonstrates remarkable data efficiency, achieving substantial accuracy gains on both in-domain and out-of-domain mobile GUI tasks using only 136 training examples. GUI-R1 [30] further boosts the real-world problem-solving capabilities of MLLMs via unified action space modeling and policy optimization, achieving state-of-the-art performance across multiple platforms in a highly data-efficient manner. InfiGUI-R1 [124] posits that advancing GUI agents requires a fundamental shift from reactive actors to deliberative reasoners and introduces the Actor2Reasoner framework, a novel two-stage training methodology. Specifically, it first injects explicit spatial reasoning capabilities through distillation and then enhances deliberation via reinforcement learning with sub-goal guidance and error recovery scenarios construction, yielding superior cross-platform performance. ComfyUI-R1 [120] presents a two-stage training framework that achieves cutting-edge automated workflow generation. The framework first adapts a model to the ComfyUI domain via CoT fine-tuning and then enhances its reasoning through RL with a novel rule-metric hybrid reward.

As a common and stable approach, offline learning from demonstration trajectories provides a solid foundation for GUI automation. However, models trained this way lack the robustness to handle real-world challenges such as unexpected events and execution errors. To bridge this gap,

research has pivoted to training models via direct online interaction in dynamic GUI environments.

Learning from online GUI Environments: GUI-Reflection [121] significantly enhances the self-reflection and error recovery capabilities of GUI automation by introducing automated data generation and iterative online tuning. This creates a new paradigm for building robust GUI agents capable of autonomous operation and error correction without the need for human annotation. ZeroGUI [123] also introduces a scalable online learning framework that eliminates the dependency on human annotations by automating both task generation and reward estimation through MLLMs. Leveraging the tailored two-stage RL process, the GUI agent enables continuous adaptation to dynamic GUI environments via autonomous interaction and self-improvement. WebAgent-R1 [122] presents an end-to-end RL framework that addresses the challenges of multi-turn decision-making in dynamic web environments by learning directly from binary task-completion rewards. UI-TARS [125] is a novel end-to-end native GUI agent that achieves unprecedented performance by integrating four key innovations: enhanced perception with large-scale GUI data, unified cross-platform action modeling, deliberate System-2 reasoning, and iterative self-improvement through reflective online trace tuning. UI-TARS-2 [308] further features a next-generation native GUI agent. Through a systematic methodology that incorporates scalable data generation, stabilized multi-turn reinforcement learning, hybrid environment integration, and a unified sandbox platform, it achieves state-of-the-art performance on both standard GUI benchmarks and complex game environments.

4.3.2 Agentic Physical Environment Interaction

Embodied AI distinguishes itself by creating autonomous agents capable of active perception, deliberate reasoning, and physical interaction within real-world environments. This paradigm aligns closely with agentic MLLMs, as both transcend passive comprehension to exhibit goal-driven, intentional behavior. By integrating sensing, planning, and acting in a closed-loop system, embodied agents underscore a pivotal shift toward models that not only interpret context but also engage with it dynamically. In this section, we explore the core capabilities that enable autonomous operation and structure our discussion into four key areas: **embodied perception, planning, navigation and manipulation**.

Embodied Perception: A substantial body of research is dedicated to embodied perception, a foundational concept in embodied AI wherein an agent acquires information through active, deliberate environmental exploration to guide its subsequent actions [328, 329, 330]. For instance, ALP [126] proposes an embodied learning framework that integrates action-aware representation learning with active environmental exploration to learn more robust and generalizable visual representations compared to static dataset training approaches. EAR [127] proposes an uncertainty-aware active recognition framework that models visual exploration as sequential evidence gathering with theoretical uncertainty quantification and reliable prediction. Incentivized by a tailored open-world reward function, this framework demonstrates superior performance in both recognition accuracy and robustness.

Embodied Planning: Building upon the actively perceptual understanding, embodied planning requires the agent to formulate a sequence of actionable steps or decisions to achieve a long-horizon goal, effectively bridging perception with concrete execution. To advance embodied planning, Wu et al. [128] propose a novel reinforcement fine-tuning framework that integrates R1-style reasoning with structured decision-making priors. Through SFT and rule-based generalized reinforced preference optimization, this method significantly enhances embodied planning performance and generalization in interactive environments. Embodied Planner-R1 [129] also incorporates RL into planning. Leveraging sparse outcome rewards and interactive policy optimization, it demonstrates superior completion ratios and robustness across multiple benchmarks.

Embodied Navigation: As a core instantiation of embodied planning, embodied navigation focuses on the agent’s ability to traverse through physical or simulated spaces by leveraging its perceptual inputs and planned path to reach a specified destination. Towards the goal of generalist navigation agents, OctoNav [130] unifies multiple navigation tasks with a new benchmark (OctoNav-Bench) and method (OctoNav-R1). Leveraging a hybrid training paradigm, OctoNav-R1 operates in a “think-before-act” mode, demonstrating impressive navigation performance. VLN-R1 [131] introduces an end-to-end framework that enables continuous vision-language navigation through direct egocentric video-to-action translation, combining an innovative long-short memory approach and time-decayed reward mechanisms to achieve strong benchmark performance through data-efficient reinforcement learning. Nav-R1 [132] further advances embodied navigation with its Fast-in-Slow reasoning framework. This dual system separates high-level semantic planning from time-critical reactive control, enabling robust and coherent navigation in dynamic environments without sacrificing real-time results.

Embodied Manipulation: Extending beyond navigation, embodied manipulation involves the agent interacting with and altering its environment through physical actions, thereby completing embodied tasks that require both motion and interaction with objects. Specifically, VLP [133] addresses the annotation bottleneck in preference-based RL via a well-designed vision-language framework that autonomously generates language-conditioned preferences for embodied manipulation tasks, facilitating scalable policy learning and robust generalization to novel instructions and tasks. ManipLVM-R1 [134] eliminates human annotation dependency through a reinforcement learning framework with two specialized rewards: Affordance Perception Reward for spatial interaction and Trajectory Match Reward for physical plausibility. Experiments show it achieves higher performance gains and better generalization with reduced training data. Embodied-R1 [135] addresses the challenging “seeing-to-doing” gap in robotics by introducing pointing as a unified intermediate representation. Through a two-stage reinforced fine-tuning framework, it achieves exceptional zero-shot generalization, offering valuable insights for the broader embodied AI community.

5 TRAINING & EVALUATION

In order to develop and assess agentic MLLMs, three core components are indispensable: **training frameworks** that provide the algorithmic and optimization infrastructure, **training datasets** that foster agentic cross-modal alignment and robust generalization, and **evaluation datasets** that measure the capabilities of agentic MLLMs. Therefore, this section surveys the landscape of open resources for agentic MLLMs across these three dimensions, helping the community to advance agentic research.

5.1 Training Framework

In this section, we review open-source training frameworks that support agentic continual pre-training, supervised fine-tuning, and reinforcement learning. These frameworks provide code implementations and advanced training optimizations that facilitate efficient development of agentic MLLMs. A summary of training framework is shown in Table 4, with corresponding links for ease of access.

Agentic CPT/SFT Frameworks. Llama-Factory [136] is an open-source, user-friendly framework that provides efficient, extensible, and unified pipelines for fine-tuning large language models across diverse tasks and settings. Mswift [137] is a versatile framework for training, aligning, and deploying large language and multi-modal models with advanced techniques. unsloth [332] is a cross-platform toolkit enabling efficient, exact-accuracy finetuning of diverse transformer models on standard NVIDIA GPUs without hardware changes. FireAct [333] provides code, prompts, and datasets for fine-tuning language agents, along with model family descriptions for research use. AgentTuning [334] introduces instruction-tuning with agent trajectories, enhancing LLMs’ agent capabilities. LM-Flow [335] is an extensible and user-friendly toolbox for efficient finetuning of large machine learning models.

Standard RL Frameworks. verl [350] is a flexible RL training library for large language models, implementing HybridFlow RLHF. TRL [336] provides a toolkit for post-training transformers via RL algorithms such as PPO and DPO. Open R1 [337] is an open reproduction of DeepSeek-R1’s reasoning pipeline, democratizing chain-of-thought training. OpenRLHF [338] offers a scalable Ray-based RLHF framework supporting PPO and GRPO. Multimodal Open R1 adds multi-modal input support to the Open R1 [337] pipeline. Logic-RL [339] introduces rule-based RL to teach logical reasoning through strict reward shaping. EasyR1 [340] is an efficient RL training framework supporting multimodality, achieving gains on reasoning benchmarks. Simple-R1 [341] explores “zero-start” RL training, showing even small models can benefit from RL on reasoning tasks. Light-R1 [342] combines supervised fine-tuning, DPO, and RL to build reasoning models from scratch. R1-V [71] improves VLM reasoning at a very low cost, demonstrating strong generalization. AReaL [343] is a fully asynchronous, open-source RL training system for large reasoning models that emphasizes reproducibility and accessibility for building AI agents.

Agentic RL Frameworks. RLFactory is an agentic RL post-training framework that decouples environment setup from training and supports asynchronous tool-calling for

TABLE 4: Summary of training framework for agentic CPT, SFT, and RL.

Framework	Link	Type	Supports MLLM	Key Features
Agentic CPT/SFT Frameworks				
LLaMA-Factory [136]	Code	Agentic CPT/SFT	Yes	Easy, Various and Efficient Fine-tuning
MS-Swift [137]	Code	Agentic CPT/SFT	Yes	Scalable Lightweight Infrastructure
Megatron-LM [331]	Code	Agentic CPT/SFT	Yes	GPU-optimized library
Unsloth [332]	Code	Agentic CPT/SFT	Yes	Accurate, Accessible, Efficient
FireAct [333]	Code	Agentic CPT/SFT	No	Language Agent Fine-tuning
AgentTuning [334]	Code	Agentic CPT/SFT	No	Generalized Agent Abilities
LMFlow [335]	Code	Agentic CPT/SFT	Yes	Extensible, Efficient, User-friendly, Open
Standard RL Frameworks				
TRL [336]	Code	RL	No	HuggingFace PPO/DPO Fine-tuning
Open R1 [337]	Code	RL	No	DeepSeek-R1 Reproduction
OpenRLHF [338]	Code	RL	No	Comprehensive, Lightweight, Easy-to-use
Multimodal Open R1	Code	RL	Yes	Multimodal R1 Training
Logic-RL [339]	Code	RL	No	Rule-based RL Reasoning
EasyR1 [340]	Code	RL	Yes	Efficient Multi-modal RL
Simple-R1 [341]	Code	RL	No	Simple RL Reasoning
Light-R1 [342]	Code	RL	No	Curriculum SFT + RL
R1-V [71]	Code	RL	Yes	General VLM RL
ARL [343]	Code	RL	No	Fully Asynchronous RL
Agentic RL Frameworks				
verl	Code	Agentic RL	Yes	Flexible, Efficient RL library
RLFactory	Code	Agentic RL	Yes	Easy, Efficient Agentic Learning
Visual-ARFT [344]	Code	Agentic RL	Yes	Flexible Agentic LLM
rLLM [138]	Code	Agentic RL	No	Customizable Agent Training
Search-R1 [345]	Code	Agentic RL	No	LLM with Search Tool
MMSearch-R1 [23]	Code	Agentic RL	Yes	Multimodal Search Agent
Agent Lightning [142]	Code	Agentic RL	No	Train-any-agent without Modifying
RAGEN [346]	Code	Agentic RL	No	RL + LLM + Agents
MARTI [347]	Code	Agentic RL	No	Multi-agent RL
MiroRL [348]	Code	Agentic RL	No	Multi-turn MCP Tool
ROLL [141]	Code	Agentic RL	No	User-friendly Large-scale RL
SkyRL [140]	Code	Agentic RL	No	Modular Full-stack RL
AWorld [349]	Code	Agentic RL	No	Agent Self-improvement at Scale
AgentFly [139]	Code	Agentic RL	Yes	Multi-turn, Async tool, Multimodal

faster agent learning. Visual-ARFT [344] equips open-source LVLMS with flexible agentic abilities for real-time web browsing and image manipulation, and introduces the MAT benchmark to evaluate multimodal search and coding skills. The rLLM framework [138] provides abstractions to define custom language agents and environments, unifying inference and training with efficient scaling. Search-R1 [345] trains LLMs to interleave reasoning with search engine calls, encouraging retrieval-based reinforcement learning. MMSearch-R1 [23] enables multi-modal models to perform multi-turn real-world search with reinforcement. Agent Lightning [142] can train virtually any agent with RL while requiring minimal code changes. RAGEN [346] leverages RL to train LLM-based reasoning agents in stochastic environments, enabling self-evolution behaviors. MARTI [347] combines centralized multi-agent interactions with distributed training, supporting scalable LLM collaboration. MiroRL [348] is the first RL framework enabling multi-turn MCP tool calls, offering agents seamless access to diverse tools while ensuring stable, efficient, and scalable training. ROLL [141] is a unified and user-friendly RL library for large-scale LLM optimization. SkyRL [140] is a modular full-stack RL library that integrates agent layers, training modules, and environments for multi-turn tasks. AWorld [349] enables large-scale agent self-improvement through continual learning from knowledge and experience. AgentFly [139] is an extensible RL framework for multi-turn, asynchronous, and multimodal agent training with easy tool and reward integration.

5.2 Training dataset

In this section, we review publicly available training datasets that support the development of agentic capabilities, including internal intelligence, external tool invocation,

and environment interaction. Corresponding links are provided for easy access and practical use, as shown in Table 5.

Agentic Internal Intelligence Datasets. We summarize the training datasets that aim to enhance agentic internal intelligence capabilities, namely reasoning, reflection, and memory. MAVIS [143] constructs valuable mathematical visual reasoning rationales through automated generation. R³V [79] provides 5K response-wise reflection SFT samples annotated by GPT. LLaVA-CoT [144] offers 100K structured chain-of-thought SFT samples distilled from GPT-4o. Mulberry-260K [19] leverages collective MCTS to search 260K reasoning and reflection data. Vision-R1 [15] utilizes GPT to generate cold-start data for RL, which contains a substantial amount of reflective content. R1-Onevision [351] also generates cold-start thinking data from complex visual reasoning tasks. MMK12 [67] collects new mathematics problems from textbooks and examination papers ranging from elementary to high school levels. OpenVLThinker [70] provides cold-start SFT data and RL data for curriculum-based reinforcement learning. ThinkLite-VL [65] repurposes MCTS to identify hard sample for effective RL optimization. Revisual [352] comprises 47K textual thought samples with reasoning paths, augmented by 31K text and 21K multimodal questions for RL. GThinker [80] adopts an iterative annotation process to generate 7K reasoning paths for SFT, followed by 4K curated samples for RL. Video-R1 [73] constructs 165K cold-start SFT samples and 260K RL training samples, both comprising image and video data. MedTVT-QA [168] is a curated instruction dataset featuring question-answer pairs for physiological interpretation and disease diagnosis using a chain-of-evidence approach. WeThink [69] introduces a scalable pipeline that generates context-aware, reasoning-centric QA pairs from images, yielding 120K multimodal QA pairs with annotated reasoning paths. AVQA-R1-6K [247] is a multimodal dataset of

TABLE 5: Summary of datasets for **training** agentic MLLMs, where T, I, V, and A represent text, image, video, and audio.

Training Dataset	Link	Stage	Type	Scope	Modality	Samples
Agentic Internal Intelligence						
MAVIS [143]	Data	SFT	Reasoning	Math	T, I	834K
R3V [79]	Data	SFT	Reasoning, Reflection	Chart, Math	T, I	5K
LLaVA-CoT-100k [144]	Data	SFT	Reasoning	Diverse	T, I	100K
Mulberry-260K [19]	Data	SFT	Reasoning, Reflection	Diverse	T, I	260K
Vision-R1-cold-200K [15]	Data	SFT	Reasoning, Reflection	Diverse	T, I	200K
R1-OneVision [351]	Data	SFT + RL	Reasoning	Diverse	T, I	155K
MM-K12 [67]	Data	RL	Reasoning	Math	T, I	15K
OpenVLThinker [70]	Data	SFT + RL	Reasoning	Diverse	T, I	12K
ThinkLite-VL [65]	Data	RL	Reasoning	Diverse	T, I	11K
Revisual-R1 [352]	Data	SFT + RL	Reasoning	Diverse	T, I	99K
GThinker-11k [80]	Data	SFT + RL	Reasoning	Diverse	T, I	11K
Video-R1 [73]	Data	SFT + RL	Reasoning	Diverse	T, I, V	425K
MedTVT-QA [168]	Data	SFT + RL	Reasoning	Medical	T, I	8K
WeThink [69]	Data	SFT + RL	Reasoning	Diverse	T, I	120K
AVQA-R1-6K [247]	Data	RL	Reasoning	Diverse	T, I, A	6K
Video-XL-pro [353]	Data	SFT	Memory	Diverse	T, V	3,000K
Long-VILA [89]	Data	SFT + RL	Memory	Diverse	T, V	71K
Agentic External Tool Invocation						
Search-R1 [101]	Data	RL	Search	Multi-hop	T	170K
Search-o1 [102]	Data	RL	Search	Multi-hop	T	1K
R1-Searcher [354]	Data	RL	Search	Multi-hop	T	8K
FVQA [23]	Data	RL	Search	Multi-hop	T, I	5K
MAT-Training [26]	Data	RL	Search, Code	Multi-hop, Code	T, I	3K
MathCoder [104]	Data	SFT	Code	Math, Code	T	80K
ReTool [25]	Data	SFT	Code	Math, Code	T	2K
ToRL [106]	Data	RL	Code	Math, Code	T	28K
rStar-Coder [355]	Data	SFT + RL	Code	Math, Code	T	580K
DeepEyes [27]	Data	RL	Visual Processing	Diverse	T, I	47K
Pixel-Reasoner [111]	Data	SFT + RL	Visual Processing	Diverse	T, I	23K
Chain-of-Focus [356]	Data	SFT	Visual Processing	Diverse	T, I	5K
Mini-o3 [28]	Data	SFT + RL	Visual Processing	Diverse	T, I	14K
Thyme [114]	Data	SFT + RL	Visual Processing	Diverse	T, I	401K
Agentic Environment Interaction						
GUI-World [145]	Data	SFT	Virtual	GUI	T, V	12K
Show-UI [31]	Data	SFT	Virtual	GUI	T, I	8K
GUI-R1-3K [30]	Data	RL	Virtual	GUI	T, I	3K
UI-R1 [173]	Data	RL	Virtual	GUI	T, I	136
GUI-Reflection [121]	Data	SFT	Virtual	GUI	T, I	296K
VLN-Ego [131]	Data	SFT + RL	Physical	Navigation	T, V	1.8M
InternData-N1 [146]	Data	SFT	Physical	Navigation	T, V	370K
VLA-IT [357]	Data	SFT	Physical	Manipulation	T, I	650K

synchronized audio-image pairs with multiple-choice questions. Long-VILA [89] and Video-XL-pro [353] introduce extended long-form video datasets for vision-language fine-tuning and enhance memory modeling.

Agentic External Tool Invocation. We summarize the training datasets for agentic external tool invocation, covering tasks such as search, code, and visual processing. Search-R1 [101], Search-o1 [102], and R1-Search [354] contribute text-based reinforcement learning datasets tailored to knowledge-intensive search. Subsequently, FVQA [23] and MAT [26] introduce knowledge-intensive multimodal datasets. These knowledge-intensive, multi-hop datasets are built from challenging and up-to-date knowledge transformed into QA pairs, as exemplified by methods such as WebSailor [235], WebDancer [358], and AgentFounder [227]. MathCoder [104] and ReTool [25] provide code datasets for SFT, while ToRL [106] and rStar-Coder [355] construct datasets suitable for reinforcement learning in agentic training. Besides, several projects such as DeepEyes [27], Pixel-Reasoner [111], and Thyme [114], have open-sourced curated datasets for interleaved text-and-image reasoning, which can be used for SFT or RL training.

Agentic Environment Interaction. We summarize the training datasets for agentic environment interaction, spanning both virtual and physical environments. Specifically, GUI-World [145] introduces the first video-based GUI dataset, while Show-UI [31], GUI-R1 [30], UI-R1 [173] and GUI-Reflection [121] provide high-quality image-based alternatives. In embodied AI, the navigation domain is supported by datasets such as VLN-Ego [131] and InternData-

N1 [146], whereas the VLA-IT [357] dataset serves as a key resource for embodied manipulation.

5.3 Evaluation Dataset

We survey the benchmarks used to evaluate the agentic capabilities of MLLMs, as presented in Table 6.

5.3.1 Benchmark Internal Intelligence

- **Benchmark Reasoning and Reflection Capabilities.** (1) General Problems. We review recent benchmarks for general visual question answering that are relatively more challenging and require reasoning, including MMBench v1.1 [359], M3CoT [269], MME-CoT [240], and MMMU-Pro [148]. (2) STEM Problems. Science, technology, engineering, and mathematics (STEM) problems are more challenging and complex, requiring MLLMs to possess stronger long-chain reasoning and reflective capabilities in order to solve them effectively, including MathVision [245], MathVerse [239], OlympiadBench [149], MMReason [241], WeMath [362], and VideoMathQA [363]. (3) Chart and Document Problems. Chart and document problems require cross-modal alignment and numerical reasoning, as illustrated by benchmarks such as CharXiv [364] and MMLongBench-Doc [150].
- **Benchmark Memory Capabilities.** Evaluating the memory capabilities of MLLMs focuses on their ability to retain and utilize information over long multi-modal contexts and multi-turn conversations. Benchmarks in

TABLE 6: Summary of datasets for **evaluating** agentic MLLMs, where T, I, and V represent text, image, and video.

Benchmark	Link	Type	Scope	Modality	Samples
Agentic Internal Intelligence					
MMBench v1.1 [359]	Data	Reasoning	General	T, I	3,217
ZeroBench [360]	Data	Reasoning	General	T, I	100
MMMU-Pro [148]	Data	Reasoning	General	T, I	3,460
MME-CoT [240]	Data	Reasoning	General	T, I	1,130
M3CoT [269]	Data	Reasoning	General	T, I	11,459
ZebraLogic [361]	Data	Reasoning, Reflection	STEM	T, I	1,000
ZeroBench [360]	Data	Reasoning, Reflection	STEM	T, I	100
OlympiadBench [149]	Data	Reasoning, Reflection	STEM	T, I	8,476
MathVision [245]	Data	Reasoning, Reflection	STEM	T, I	3,040
MathVerse [239]	Data	Reasoning, Reflection	STEM	T, I	2,612
MMReason [241]	Data	Reasoning, Reflection	STEM	T, I	2,941
WeMath [362]	Data	Reasoning, Reflection	STEM	T, I	6,500
VideoMathQA [363]	Data	Reasoning, Reflection	STEM	T, V	2,100
CharXiv [364]	Data	Reasoning	Chart	T, I	2,323
LoCoMo [365]	Data	Memory	General	T, I	50
MileBench [151]	Data	Memory	General	T, I	6,440
MMLongBench-Doc [150]	Data	Reasoning, Memory	Doc	T, I	135
LongVideoBench [366]	Data	Reasoning, Memory	General	T, V	6,678
LVBench [367]	Data	Reasoning, Memory	General	T, V	1,549
Agentic External Tool Invocation					
Humanity’s Last Exam [368]	Data	Search	General	T, I	2,500
MM-BrowseComp [152]	Data	Search	General	T, I	224
BrowseComp-VL [24]	Data	Search	General	T, I	399
FVQA [23]	Data	Search	General	T, I	1,800
MMSearch [192]	Data	Search	General	T, I	300
MMSearch-Plus [98]	Data	Search	General	T, I	311
ViDoSeek [369]	Data	Search	Doc	T, I	1,200
MAT [26]	Data	Search, Code	General	T, I	350
WebMMU [154]	Data	Code	General	T, I	10,199
Design2Code [155]	Data	Code	Webpage	T, I	484
Flame-React-Eval [370]	Data	Code	UI	T, I	80
V*Bench [371]	Data	Visual Processing	General	T, I	191
HRBench [372]	Data	Visual Processing	General	T, I	200
Agentic Environment Interaction					
ScreenSpot [373]	Data	Virtual	General	T, I	1200
ScreenSpot-Pro [374]	Data	Virtual	General	T, I	1,581
AndriodWorld [246]	Data	Virtual	Andriod	T, I	116
AndriodControl [375]	Data	Virtual	Andriod	T, I	15,283
OSWorld [157]	Data	Virtual	Computer	T, I	369
WebWalkerQA [233]	Data	Virtual	Web	T, I	680
OmniACT [376]	Data	Virtual	Web,Desktop	T, I	9,802
LH-VLN [159]	Data	Physical	Navigation	T, I	3,260
HA-VLN [377]	Data	Physical	Navigation	T, I	16,844
VLABench [158]	Data	Physical	Manipulation	T, I	2,164

this category include MileBench [151], MMLongBench-Doc [150], LongVideoBench [366], and LVBench [367], which assess how well models can preserve contextual information, recall relevant details, and maintain coherent reasoning across extended interactions.

5.3.2 Benchmark External Tool Invocation.

- **Benchmark Search Capabilities.** Evaluating agentic search capabilities typically relies on benchmarks composed of multi-hop, knowledge-intensive, and up-to-date questions. Such tasks require the model not only to retrieve relevant information from external resources but also to integrate evidence across multiple sources and reason over them to reach a correct conclusion. Representative benchmarks include Humanity’s Last Exam [368], MM-BrowseComp [152], BrowseComp-VL [24], FVQA [23], MMSearch [192], MMSearch-Plus [98], ViDoSeek [369], and MAT [26].
- **Benchmark Code Capabilities.** Code benchmarks evaluate how well MLLMs can generate code across multiple languages, *e.g.*, Python, JavaScript, and SQL. Representative benchmarks include WebMMU [154], Design2Code [155] and Flame-React-Eval [370]. Additionally, several advanced mathematical benchmarks, such as AIME2024 and AIME2025, are commonly employed to evaluate the code-integrated reasoning capabilities.
- **Benchmark Visual Processing Capabilities.** Benchmarks for high-resolution image understanding (*e.g.*, V*Bench [371] and HRBench [372]) evaluate the visual

processing capability that requires agentic MLLMs to perform operations like cropping and zooming to uncover visual clues, leading to enhanced image comprehension.

5.3.3 Benchmark Environment Interaction

- **Benchmark Virtual Interaction Capabilities.** A range of GUI benchmarks, such as ScreenSpot [373], Andriod-World [246] and OSWorld [157], serve to evaluate virtual interaction capabilities. These benchmarks provide diverse environments where agents must execute tasks by interacting with graphical user interfaces, testing their ability to understand screen elements and perform correct sequences of actions.
- **Benchmark Physical Interaction Capabilities.** In embodied AI and robotics, core physical interaction capabilities are evaluated across key domains, with navigation assessed on benchmarks such as LH-VLN [159] and HA-VLN [377], and manipulation evaluated using VLABench [158].

6 APPLICATION

Agentic MLLMs, endowed with strong generalization capabilities and integrated agentic functionalities, have demonstrated remarkable potential across a broad spectrum of downstream tasks. Unlike previous MLLM agents that are often restricted to specific domains, agentic MLLMs can reason, reflect, leverage memory, invoke various external tools,

and interact with dynamic environments, enabling them to handle complex real-world scenarios. This transformative paradigm has garnered growing attention from diverse research communities, offering fresh insights into long-standing challenges and unlocking new opportunities for practical applications in areas such as Deep Research, Embodied AI, Healthcare, GUI Agents, Autonomous Driving, and Recommender Systems. In the following subsections, we present an overview of these representative applications and highlight how agentic MLLMs are reshaping them.

6.1 Deep Research

Deep Research (DR) represents a milestone in agentic intelligence, showcasing the ability of MLLMs to autonomously conduct multi-step, goal-directed research for high-intensity knowledge work. Unlike conventional models that rely on single-turn retrieval or user-driven prompting, Deep Research integrates multi-step reasoning and tool use to automate information discovery and synthesis, thereby assisting domains such as finance, science, policy, and education in handling complex tasks [160, 378, 379, 380, 381]. Recently, a variety of Deep Research agents have emerged, including OpenAI Deep Research [160], Gemini Deep Research [161], Grok DeepSearch [162], Perplexity Deep Research [189], Copilot Researcher [382], Kimi-Researcher [383], AutoGLM [384], Tongyi Deep Research [163], MiroThinker [188], and Manus [385]. These Deep Research systems demonstrate strong capabilities in open-ended, knowledge-intensive tasks, enabling them to tackle the kinds of complex, real-world problems that people encounter in both professional and everyday contexts. It thus marks a significant step toward practical, autonomous AI systems capable of scalable and verifiable research.

6.2 Embodied AI

Embodied AI marks a transformative shift from passive perception to active engagement in physical environments, with vision-language-action (VLA) models emerging as a pivotal architectural framework [32, 166, 386, 387]. These models integrate multimodal reasoning with motion control to translate high-level linguistic and visual inputs into executable action sequences, thereby serving as the cognitive core for next-generation robotic systems. In robotics, VLA-powered agents demonstrate remarkable open-world generalization [3, 388, 389, 390, 391], significantly expanding the scope of complex tasks achievable by machines. Beyond technical advancement, this synergy drives substantial commercial value across logistics, smart manufacturing, and personalized service domains, offering scalable, intelligent solutions for dynamic real-world applications [392, 393].

6.3 Healthcare

The rapid advancement of MLLMs has spurred growing interest in their application within healthcare contexts. Unlike general domains, medicine requires exceptional reliability, strict control of hallucinations, and robust interpretability. Early approaches such as LLaVA-Med [394] and HuatuoGPT series [395, 396, 397] rely on SFT with curated medical QA data, but often exhibit limited generalization. Subsequent efforts like HuatuoGPT-o1 incorporate

RL (e.g., PPO) to activate reasoning and self-reflection, markedly improving diagnostic accuracy [167, 168, 398]. Beyond enhancing intrinsic model capabilities, systems such as MMed-RAG [399] and MedResearcher-R1 [400] further integrate external tools like domain-aware retrieval and medical knowledge graphs. These agentic MLLMs combine sophisticated retrieval mechanisms or other advanced tools with RL to achieve state-of-the-art performance on complex medical reasoning tasks [401, 402, 403]. Moreover, medical embodied AI systems [156], such as those in surgical robotics [404, 405, 406], are also showing promising application prospects and practical value, further extending the impact of agentic MLLMs into physical clinical interventions.

6.4 GUI Agents

GUI agents represent a breakthrough application of agentic MLLMs, fundamentally reshaping human-computer interaction [30, 125, 173, 308, 407]. They demonstrate remarkable capability in automating complex digital tasks across diverse software environments and operating systems, including web scenarios [122], mobile platforms [407], and desktop interfaces [157, 408]. By visually perceiving the screen, comprehending natural language commands, and executing precise low-level actions (e.g., clicks, typing, scrolling), they enable a wide range of sophisticated applications, including fundamental tasks like file management and web operations [409], and more advanced capabilities from cross-app workflow orchestration [410] to personalized user support [411]. The advancement of GUI agents holds significant potential to enhance digital accessibility and operational efficiency, thereby offering substantial benefits to both commercial ecosystems and broader societal infrastructures.

6.5 Autonomous Driving

The application of agentic MLLMs in autonomous driving represents a rapidly evolving research frontier aimed at enhancing complex decision-making and interaction capabilities [174, 175, 176, 412]. One line of work incorporates CoT reasoning into autonomous driving systems, utilizing the sophisticated cognitive capabilities of MLLMs to generate accurate and interpretable motion trajectories [174, 175, 412, 413, 414]. Another category of methods integrates external tools, such as object detection, depth estimation, and occupancy prediction, to enhance perceptual robustness and situational awareness. Through SFT combined with RL training, such models learn to autonomously invoke and leverage these tools, significantly improving the robustness and generalization of driving policies in open-world scenarios [176]. Together, these efforts highlight a clear trend toward building more reliable, transparent, and tool-augmented MLLM-based agents for autonomous driving. By combining internal reasoning capability with external perceptual tools and advanced training paradigms, agentic MLLMs are poised to overcome key challenges in real-time decision-making, safety assurance, and scalable deployment in dynamic driving environments.

6.6 Recommender System

Traditional MLLM-based recommender systems [9, 415] primarily enhance existing recommendation pipelines by

leveraging multimodal representations and language understanding to improve ranking, retrieval, and conversational interactions. However, these systems typically remain reactive: they rely on pre-defined objectives (e.g., click-through rate prediction), static user profiles, and limited dialogue rounds to refine outputs. While MLLMs enable richer modeling of user intent and item semantics, they still lack deeper autonomy and adaptability. Recently, agentic MLLM recommender systems (MLLM-ARS) [178, 179, 180, 181, 416] have emerged to transcend this paradigm by embedding reasoning, reflection, memory, tool use, and virtual interaction within the recommendation process. Rather than passively responding to user requests, agentic recommenders proactively explore user preferences, simulate future behaviors, and adapt strategies over time. They integrate multimodal cues with agentic capabilities such as reasoning, reflection, and role-playing to deliver interactive, context-aware, and personalized experiences. Crucially, these systems evolve dynamically, balancing immediate feedback with long-horizon personalization, paving the way for recommender systems that are not only responsive but also autonomous, transparent, and continuously self-improving.

7 CHALLENGES AND FUTURE DIRECTIONS

Despite recent progress, the development of agentic MLLMs is still in its early stages, and many challenges remain to be addressed. This section discusses these limitations and outlines potential directions for future research.

7.1 Richer Action Space of Agentic MLLM

Agentic MLLMs have demonstrated remarkable capabilities in handling complex tasks. However, the action space of existing models, and the range of tools they can access is often restricted to a single type [23, 25]. Recent studies have integrated a wider range of tool usage. For example, Visual-ARFT [26] can perform both search and code execution, while WebWatcher [24] supports even richer functionalities, including search, code interpretation, and internal OCR. Looking ahead, future agentic MLLMs are expected to operate with a richer action space, equipped to invoke a broader spectrum of external tools and services. They may seamlessly integrate with data analysis platforms, simulation environments, multimodal sensors, and interactive APIs, enabling more adaptive and generalizable agentic behaviors across diverse real-world scenarios.

7.2 Efficient Agentic MLLMs

While agentic MLLMs excel at handling complex problems through multi-turn reasoning and external tool invocation, these iterative processes substantially increase their computational and reasoning overhead. In some cases, models may require up to thirty minutes to complete a single task [160], imposing significant costs on both training and inference. Such inefficiency poses challenges for real-time applications and large-scale deployment, where latency, energy consumption, and resource constraints become critical considerations. Although some studies have accelerated long-chain reasoning [417, 418, 419, 420], research on speeding up tool invocation remains limited. To address these issues, future

research should focus on improving the efficiency of agentic MLLMs, accelerating both training and inference without compromising performance. By enhancing computational efficiency, agentic MLLMs can move closer to practical, scalable deployment across diverse real-world environments.

7.3 Long-term Agentic Memory

Long-term memory allows agentic MLLMs to plan, reason, and interact in ways that support continuity, adaptation, and long-term experience accumulation over time. Although recent studies have explored agentic memory [21, 22, 88, 93], most of these works have focused primarily on the language modality, with limited exploration of multimodal settings. At the same time, the effective length of memory in current systems remains highly constrained, restricting their ability to sustain coherent knowledge across longer time horizons. Future work should design persistent memory architectures that allow models to accumulate, organize, and retrieve knowledge across extended time spans. Such memory must be both scalable, capable of processing the vast multimodal streams agents encounter, and selective, able to filter, compress, and prioritize experiences relevant for reasoning, ultimately supporting evolving memory systems that foster personalization, sustained collaboration, and adaptive problem-solving. Ultimately, long-term agentic memory is not just a technical refinement but a prerequisite for creating enduring partners capable of continuous learning and alignment with human goals.

7.4 Agentic Training and Evaluation Dataset

Currently, the development of agentic MLLMs is still at a very early and exploratory stage, and one of the most pressing challenges lies in the scarcity of training datasets specifically designed for agentic behaviors. Tongyi Lab [227, 233, 236] proposes a fully automated pipeline for generating synthetic agentic trajectories, supporting CPT, SFT, and RL. However, much of this data remains inaccessible to the research community and lacks sufficient exploration in multimodal domains. Therefore, an urgent research direction lies in developing effective and efficient methods for synthesizing high-quality multimodal agentic trajectory data. In addition, to evaluate the performance of agentic MLLMs, several recent benchmarks have been established, such as MM-BrowseComp [152] and BrowseComp-VL [24]. However, these benchmarks primarily focus on specific aspects of agentic behavior, while certain actions, such as memory utilization and the ability to coordinate reasoning across multiple tool invocations, still lack effective evaluation datasets. Moreover, robust methods for assessing whether actions are correctly executed remain underexplored.

7.5 Safe Agentic MLLMs

AI safety has long been recognized as a central challenge, and prior work [421, 422, 423, 424, 425] has focused extensively on building systems that are safe and controllable. As agentic MLLMs become increasingly autonomous in planning, tool invocation, and environment interaction, ensuring their safety will be a critical research priority. Unlike static models, agentic systems dynamically generate

action sequences that may call external tools, APIs, or even physical devices, thereby amplifying the risks of unintended consequences [426]. For instance, a model conducting web search may retrieve incorrect or harmful information, which can bias subsequent MDP-based decision making and lead to unsafe downstream actions. In multimodal settings, the difficulty is further magnified, as ambiguous or adversarial inputs can propagate across modalities and destabilize agent behavior. Addressing these challenges requires a combination of rigorous benchmarking, adversarial stress-testing, and the integration of normative frameworks, ultimately ensuring that agentic MLLMs remain reliable, controllable, and aligned with human intent as they advance toward more general autonomy.

8 CONCLUSION

This survey charts the recent advances of agentic MLLMs, marking a pivotal shift from traditional MLLM agents to models with agentic capabilities. We begin by discussing MLLM agents and agentic MLLMs, the latter distinguished by dynamic workflows, proactive execution of actions, and strong generalization across domains. We then introduce agentic foundational MLLMs, action space, CPT, SFT, RL, and evaluation methodologies, which together serve as the preliminary knowledge base. Building on it, we propose a threefold taxonomy that organizes MLLM agentic capabilities into: (i) internal intelligence, where reasoning, reflection, and memory coordinate long-horizon decisions; (ii) external tool invocation, where models proactively call search engines, code executors, and visual processing to acquire and manipulate information; and (iii) environment interaction, where agents act within virtual and physical settings to obtain feedback and continuously refine their plans through iterative adaptation. In addition, we consolidated open-source training frameworks, training datasets, and evaluation benchmarks to provide a practical reference that can ground and accelerate future research, and we summarized emerging agentic applications across diverse scenarios. We also track notable developments through a real-time [GitHub repository](#) and hope that these resources will help accelerate the advancement of agentic MLLMs.

REFERENCES

- [1] OpenAI. Gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024.
- [2] S. Bai et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] G. R. Team et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.
- [4] J. Xu et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.
- [5] Y. Yao et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [6] S. Wu et al. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.
- [7] J. Zhan et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.
- [8] A. Team. Claude 4 sonnet, February 2025.
- [9] Y. Ye et al. Harnessing multimodal large language models for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13069–13077, 2025.
- [10] W. Wu et al. Gpt4vis: What can gpt-4 do for zero-shot visual recognition? *arXiv preprint arXiv:2311.15732*, 2023.
- [11] Z. Yang et al. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*, 2023.
- [12] X. Li et al. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18061–18070, 2024.
- [13] W. Huang et al. Keeping yourself is important in downstream tuning multimodal large language model. *arXiv preprint arXiv:2503.04543*, 2025.
- [14] J. Zhang et al. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.
- [15] W. Huang et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- [16] J. Zhang et al. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.
- [17] Y. Wang et al. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
- [18] Q. Chen et al. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.
- [19] H. Yao et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search, 2024.
- [20] Z. Wan et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*, 2025.
- [21] W. Xu et al. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.
- [22] S. Yan et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025.
- [23] J. Wu et al. Mmsearch-r1: Incentivizing llms to search. *arXiv preprint arXiv:2506.20670*, 2025.
- [24] X. Geng et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.
- [25] J. Feng et al. Retool: Reinforcement learning for strategic tool use in llms. *arXiv preprint arXiv:2504.11536*, 2025.
- [26] Z. Liu et al. Visual agentic reinforcement fine-tuning. *arXiv preprint arXiv:2505.14246*, 2025.
- [27] Z. Zheng et al. Deepeyes: Incentivizing “thinking with images” via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.
- [28] X. Lai et al. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.
- [29] Z. Su et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.
- [30] R. Luo et al. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025.
- [31] K. Q. Lin et al. Showui: One vision-language-action model for gui visual agent. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19498–19508, 2025.
- [32] M. J. Kim et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [33] H. Chen et al. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv*

- preprint arXiv:2504.11468*, 2025.
- [34] J. Xie et al. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024.
- [35] L. Wang et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.
- [36] T. Cao et al. Phishagent: a robust multimodal agent for phishing webpage detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27869–27877, 2025.
- [37] J. Y. Koh et al. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024.
- [38] X. Deng et al. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.
- [39] G. Verma et al. Adaptagent: Adapting multimodal web agents with few-shot learning from human demonstrations. *arXiv preprint arXiv:2411.13451*, 2024.
- [40] S. Yao et al. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.
- [41] J. Wang et al. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *Advances in Neural Information Processing Systems*, 37:2686–2710, 2024.
- [42] B. Yang et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025.
- [43] A. Dubey et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [44] J. Xu et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.
- [45] W. Yin et al. Sail-vl2 technical report, 2025.
- [46] L. Xiaomi and C. Team. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.
- [47] V. Team et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.
- [48] J. Zhu et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
- [49] A. Marafioti et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.
- [50] G. Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- [51] M. Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.
- [52] L. C. Contributors. Llava-onevision-1.5: Fully open framework for democratized multimodal training. In *arxiv*, 2025.
- [53] Y. Li et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.
- [54] X. Dong et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.
- [55] Q. Team. Qwen3-vl: Sharper vision, deeper thought, broader action, September 2025.
- [56] Z. Wu et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.
- [57] D. Guo et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- [58] W. Wang et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.
- [59] S. Team. Step-3 is large yet affordable: Model-system co-design for cost-effective decoding, 2025.
- [60] G. Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- [61] K. Team et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [62] Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, 2025.
- [63] Baidu-ERNIE-Team. Ernie 4.5 technical report, 2025.
- [64] H. Zhang et al. Mml. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024.
- [65] X. Wang et al. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.
- [66] H. Yao et al. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*, 2025.
- [67] F. Meng et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, 2025.
- [68] H. Shen et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.
- [69] J. Yang et al. Wethink: Toward general-purpose vision-language reasoning via reinforcement learning. *arXiv preprint arXiv:2506.07905*, 2025.
- [70] Y. Deng et al. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement, 2025.
- [71] L. Chen et al. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. Accessed: 2025-02-02.
- [72] J. Xia et al. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.
- [73] K. Feng et al. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.
- [74] R. Yuan et al. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*, 2025.
- [75] K. Fan et al. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. *arXiv preprint arXiv:2505.17018*, 2025.
- [76] P. Wang et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*, 2025.
- [77] Y. Chen et al. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025.
- [78] Y. Chen et al. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *arXiv preprint arXiv:2506.16141*, 2025.
- [79] K. Cheng et al. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*, 2024.
- [80] Y. Zhan et al. Gthinker: Towards general multimodal reasoning via cue-guided rethinking. *arXiv preprint arXiv:2506.01078*, 2025.
- [81] H. Wang et al. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.
- [82] S. Yang et al. Look-back: Implicit visual re-focusing in mllm reasoning. *arXiv preprint arXiv:2507.03019*, 2025.
- [83] P. Jian et al. Look again, think slowly: Enhancing visual reflection in vision-language models. *arXiv preprint arXiv:2509.12132*, 2025.
- [84] Q. Chen et al. Long-horizon visual imitation learning via

- plan and code reflection. *arXiv preprint arXiv:2509.05368*, 2025.
- [85] H. Wei and Z. Chen. Training-free reasoning and reflection in mllms. *arXiv preprint arXiv:2505.16151*, 2025.
- [86] B. He et al. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514, 2024.
- [87] E. Song et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.
- [88] Y. Ding et al. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.
- [89] Y. Chen et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024.
- [90] W. Zhong et al. Memorybank: Enhancing large language models with long-term memory, 2023.
- [91] E. Lumer et al. Memtool: Optimizing short-term memory management for dynamic tool calling in llm agent multi-turn conversations. *arXiv preprint arXiv:2507.21428*, 2025.
- [92] P. Chhikara et al. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.
- [93] Z. Zhou et al. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*, 2025.
- [94] Z. Tan et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. *arXiv preprint arXiv:2503.08026*, 2025.
- [95] Y. Wang et al. M+: Extending memoryllm with scalable long-term memory. *arXiv preprint arXiv:2502.00592*, 2025.
- [96] Q. Wang et al. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*, 2025.
- [97] H. He et al. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024.
- [98] X. Tao et al. Mmsearch-plus: A simple yet challenging benchmark for multimodal browsing agents. *arXiv preprint arXiv:2508.21475*, 2025.
- [99] Z. Xiao et al. M2io-r1: An efficient rl-enhanced reasoning framework for multimodal retrieval augmented multimodal generation. *arXiv preprint arXiv:2508.06328*, 2025.
- [100] W. Zhang et al. Patho-agenticrag: Towards multimodal agentic retrieval-augmented generation for pathology vlms via reinforcement learning. *arXiv preprint arXiv:2508.02258*, 2025.
- [101] B. Jin et al. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.
- [102] X. Li et al. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.
- [103] Z. Gou et al. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*, 2023.
- [104] K. Wang et al. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*, 2023.
- [105] X. Guan et al. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.
- [106] X. Li et al. Torl: Scaling tool-integrated rl. *arXiv preprint arXiv:2503.23383*, 2025.
- [107] C. Li et al. Cort: Code-integrated reasoning within thinking. *arXiv preprint arXiv:2506.09820*, 2025.
- [108] M. Cao et al. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025.
- [109] M. Zhu et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 2025.
- [110] X. Zhang et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025.
- [111] A. Su et al. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.
- [112] C. Jiang et al. Vlm-r³: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*, 2025.
- [113] Z. Su et al. Openthinking: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.
- [114] Y.-F. Zhang et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.
- [115] Z. Zhou et al. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025.
- [116] Y. Xu et al. Visual planning: Let’s think only with images. *arXiv preprint arXiv:2505.11409*, 2025.
- [117] Y. Xu et al. Aguis: Unified pure vision agents for autonomous gui interaction. *arXiv preprint arXiv:2412.04454*, 2024.
- [118] Y. Liu et al. Infiguiagent: A multimodal generalist gui agent with native reasoning and reflection. *arXiv preprint arXiv:2501.04575*, 2025.
- [119] B. Zhang et al. Tongui: Building generalized gui agents by learning from multimodal web tutorials. *arXiv preprint arXiv:2504.12679*, 2025.
- [120] Z. Xu et al. Comfyui-r1: Exploring reasoning models for workflow generation. *arXiv preprint arXiv:2506.09790*, 2025.
- [121] P. Wu et al. Gui-reflection: Empowering multimodal gui models with self-reflection behavior. *arXiv preprint arXiv:2506.08012*, 2025.
- [122] Z. Wei et al. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. *arXiv preprint arXiv:2505.16421*, 2025.
- [123] C. Yang et al. Zerogui: Automating online gui learning at zero human cost. *arXiv preprint arXiv:2505.23762*, 2025.
- [124] Y. Liu et al. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. *arXiv preprint arXiv:2504.14239*, 2025.
- [125] Y. Qin et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.
- [126] X. Liang et al. Alp: Action-aware embodied learning for perception. *arXiv preprint arXiv:2306.10190*, 2023.
- [127] L. Fan et al. Evidential active recognition: Intelligent and prudent open-world embodied perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16351–16361, 2024.
- [128] D. Wu et al. Reinforced reasoning for embodied planning. *arXiv preprint arXiv:2505.22050*, 2025.
- [129] Z. Fei et al. Unleashing embodied task planning ability in llms via reinforcement learning. *arXiv preprint arXiv:2506.23127*, 2025.
- [130] C. Gao et al. Octonav: Towards generalist embodied navigation. *arXiv preprint arXiv:2506.09839*, 2025.
- [131] Z. Qi et al. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221*, 2025.
- [132] Q. Liu et al. Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884*, 2025.
- [133] R. Liu et al. Vlp: Vision-language preference learning for embodied manipulation. *arXiv preprint arXiv:2502.11918*, 2025.
- [134] Z. Song et al. Manipvlm-r1: Reinforcement learning for

- reasoning in embodied manipulation with large vision-language models. *arXiv preprint arXiv:2505.16517*, 2025.
- [135] Y. Yuan et al. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. *arXiv preprint arXiv:2508.13998*, 2025.
- [136] Y. Zheng et al. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.
- [137] Y. Zhao et al. Swift: a scalable lightweight infrastructure for fine-tuning, 2024.
- [138] S. Tan et al. rllm: A framework for post-training language agents. <https://pretty-radio-b75.notion.site/rLLM-A-Framework-for-Post-Training-Language-Agents-21b81902c146819db63cd98a54ba5f31>, 2025. Notion Blog.
- [139] R. Wang et al. Agentfly: Extensible and scalable reinforcement learning for llm agents, 2025.
- [140] S. Cao et al. Skyrl-v0: Train real-world long-horizon agents via reinforcement learning, 2025.
- [141] W. Wang et al. Reinforcement learning optimization for large-scale learning: An efficient and user-friendly scaling library. *arXiv preprint arXiv:2506.06122*, 2025.
- [142] X. Luo et al. Agent lightning: Train any ai agents with reinforcement learning, 2025.
- [143] R. Zhang et al. Mavis: Mathematical visual instruction tuning. *arXiv e-prints*, pp. arXiv-2407, 2024.
- [144] G. Xu et al. Llava-cot: Let vision language models reason step-by-step, 2025.
- [145] D. Chen et al. Gui-world: A video benchmark and dataset for multimodal gui-oriented understanding. *arXiv preprint arXiv:2406.10819*, 2024.
- [146] I.-N. D. contributors. Interndata-n1 dataset. <https://huggingface.co/datasets/InternRobotics/InternData-N1>, 2025.
- [147] Y. Wanyan et al. Look before you leap: A gui-critic-r1 model for pre-operative error diagnosis in gui automation. *arXiv preprint arXiv:2506.04614*, 2025.
- [148] X. Yue et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.
- [149] C. He et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.
- [150] Y. Ma et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.
- [151] D. Song et al. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.
- [152] S. Li et al. Mm-browsecomp: A comprehensive benchmark for multimodal browsing agents. *arXiv preprint arXiv:2508.13186*, 2025.
- [153] J. Wei et al. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.
- [154] R. Awal et al. Webmmu: A benchmark for multimodal multilingual website understanding and code generation. *arXiv preprint arXiv:2508.16763*, 2025.
- [155] C. Si et al. Design2code: Benchmarking multimodal code generation for automated front-end engineering. *arXiv preprint arXiv:2403.03163*, 2024.
- [156] Y. Liu et al. From screens to scenes: A survey of embodied ai in healthcare. *Information Fusion*, 119:103033, 2025.
- [157] T. Xie et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.
- [158] S. Zhang et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *arXiv preprint arXiv:2412.18194*, 2024.
- [159] X. Song et al. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12078–12088, 2025.
- [160] OpenAI. Introducing deep research. <https://openai.com/index/introducing-deep-research/>, 2025.
- [161] Gemini Team. Introducing gemini deep research. <https://gemini.google/overview/deep-research/>, 2025.
- [162] xAI Team. Introducing grok deepsearch. <https://x.ai/news/grok-3>, 2025.
- [163] Tongyi DeepResearch Team. Tongyi-deeprersearch. <https://github.com/Alibaba-NLP/DeepResearch>, 2025.
- [164] Y. Liu et al. Omnivia: Embodied versatile planner via task-adaptive 3d-grounded and embodiment-aware reasoning. *arXiv preprint arXiv:2509.09332*, 2025.
- [165] A. Zhai et al. Igniting vlms toward the embodied space. *arXiv preprint arXiv:2509.11766*, 2025.
- [166] D. Qu et al. Embodiedonevision: Interleaved vision-text-action pretraining for general robot control. *arXiv preprint arXiv:2508.21112*, 2025.
- [167] P. Hao et al. Surgery-r1: Advancing surgical-vqla with reasoning multimodal large language model via reinforcement learning. *arXiv preprint arXiv:2506.19469*, 2025.
- [168] Y. Zhang et al. Medtvt-r1: A multimodal llm empowering medical reasoning and diagnosis. *arXiv preprint arXiv:2506.18512*, 2025.
- [169] M. R. Rezaei et al. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between llms and evolving medical knowledge. *arXiv preprint arXiv:2502.13010*, 2025.
- [170] Z. Lei et al. Surgvisagent: Multimodal agentic model for versatile surgical visual enhancement. *arXiv preprint arXiv:2507.02252*, 2025.
- [171] Y. Xu et al. Mobilerl: Online agentic reinforcement learning for mobile gui agents, 2025.
- [172] L. Lin et al. Inframind: A novel exploration-based gui agentic framework for mission-critical industrial management. *arXiv preprint arXiv:2509.13704*, 2025.
- [173] Z. Lu et al. Ui-r1: Enhancing action prediction of gui agents by reinforcement learning. *arXiv preprint arXiv:2503.21620*, 2025.
- [174] Y. Li et al. Drive-r1: Bridging reasoning and planning in vlms for autonomous driving with reinforcement learning. *arXiv preprint arXiv:2506.18234*, 2025.
- [175] B. Jiang et al. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*, 2025.
- [176] K. Qian et al. Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. *arXiv preprint arXiv:2505.15298*, 2025.
- [177] X. Hou et al. Driveagent: Multi-agent structured reasoning with llm and multimodal sensor fusion for autonomous driving. *arXiv preprint arXiv:2505.02123*, 2025.
- [178] C. Huang et al. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734*, 2025.
- [179] F. Liu et al. Recoworld: Building simulated environments for agentic recommender systems. *arXiv preprint arXiv:2509.10397*, 2025.
- [180] S. Chen et al. Vragent-r1: Boosting video recommendation with mllm-based agents via reinforcement learning. *arXiv preprint arXiv:2507.02626*, 2025.
- [181] Y. Zhang et al. Reasonrec: A reasoning-augmented multimodal agent for unified recommendation. In *ICML 2025 Workshop on Programmatic Representations for Agent Learning*.

- [182] A. Jaech et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [183] Z. Ke et al. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*, 2025.
- [184] D. Guo et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [185] Z. Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [186] G. Zhang et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025.
- [187] K. Zhang et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.
- [188] MiroMind AI Team. Miroflow: An open-source agentic framework for deep research. <https://github.com/MiroMindAI/MiroFlow>, 2025.
- [189] Perplexity Team. Introducing perplexity deep research. <https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>, 2025.
- [190] Z. Li et al. Autoflow: Automated workflow generation for large language model agents. *arXiv preprint arXiv:2407.12821*, 2024.
- [191] X. Wang et al. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.
- [192] D. Jiang et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.
- [193] Z. Gu et al. Agentgroupchat-v2: Divide-and-conquer is what llm-based multi-agent system need. *arXiv preprint arXiv:2506.15451*, 2025.
- [194] J. Li et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.
- [195] J.-B. Alayrac et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- [196] R. Luo et al. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.
- [197] H. Liu et al. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [198] H. Liu et al. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.
- [199] A. Hurst et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [200] Y. Li et al. Mini-gemini: Mining the potential of multi-modality vision language models. *CoRR*, 2024.
- [201] B. Lin et al. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.
- [202] B. Li et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [203] P. Tong et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.
- [204] H. Laurençon et al. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.
- [205] J. Ye et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.
- [206] S. Chen et al. Ocean-ocr: Towards general ocr application via a vision-language model. *arXiv preprint arXiv:2501.15558*, 2025.
- [207] A. Hu et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.
- [208] L. Zhang et al. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*, 2024.
- [209] Y. Feng et al. Dive into moe: Diversity-enhanced reconstruction of large language models from dense into mixture-of-experts. *arXiv preprint arXiv:2506.09351*, 2025.
- [210] S. Wang et al. Scaling laws across model architectures: A comparative analysis of dense and moe models in large language models. *arXiv preprint arXiv:2410.05661*, 2024.
- [211] Z. Wu et al. Valley2: Exploring multimodal models with scalable vision-language design. *arXiv preprint arXiv:2501.05901*, 2025.
- [212] A. Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [213] A. Zeng et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.
- [214] L. Wiedmann et al. Finevision: Open data is all you need, September 2025.
- [215] W. Wang et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.
- [216] M. Zhang et al. Automated multi-level preference for mllms. *Advances in Neural Information Processing Systems*, 37:26171–26194, 2024.
- [217] W. Wang et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.
- [218] W. Huang et al. Be confident: Uncovering overfitting in mllm multi-task tuning. In *ICML*, 2025.
- [219] B. Lin et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.
- [220] Y. Li et al. Pace: Unified multi-modal dialogue pre-training with progressive and compositional experts. *arXiv preprint arXiv:2305.14839*, 2023.
- [221] J. Chen et al. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1110–1119, 2024.
- [222] Y. Li et al. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [223] T. Huai et al. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19608–19617, 2025.
- [224] S. Agarwal et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.
- [225] M. Wang et al. Two experts are all you need for steering thinking: Reinforcing cognitive effort in moe reasoning models without additional training. *arXiv preprint arXiv:2505.14681*, 2025.
- [226] D. Dai et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.
- [227] L. Su et al. Scaling agents via continual pre-training. *arXiv preprint arXiv:2509.13310*, 2025.
- [228] Z. Ke et al. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*, 2023.
- [229] Y. Chen et al. Comp: Continual multimodal pre-training for vision foundation models. *arXiv preprint arXiv:2503.18931*, 2025.
- [230] K. Gupta et al. Continual pre-training of large language models: How to (re) warm your model? *arXiv preprint*

- arXiv:2308.04014*, 2023.
- [231] Ç. Yıldız et al. Investigating continual pretraining in large language models: Insights and implications. *arXiv preprint arXiv:2402.17400*, 2024.
 - [232] W. Wu et al. Masksearch: A universal pre-training framework to enhance agentic search capability. *arXiv preprint arXiv:2505.20285*, 2025.
 - [233] J. Wu et al. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*, 2025.
 - [234] K. Li et al. Websailor-v2: Bridging the chasm to proprietary agents via synthetic data and scalable reinforcement learning. *arXiv preprint arXiv:2509.13305*, 2025.
 - [235] K. Li et al. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*, 2025.
 - [236] Z. Tao et al. Webshaper: Agentially data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*, 2025.
 - [237] S. Yao et al. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
 - [238] J. Schulman et al. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
 - [239] R. Zhang et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.
 - [240] D. Jiang et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.
 - [241] H. Yao et al. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward agi. *arXiv preprint arXiv:2506.23563*, 2025.
 - [242] S. Huang et al. Planning, creation, usage: Benchmarking llms for comprehensive tool utilization in real-world complex scenarios. *arXiv preprint arXiv:2401.17167*, 2024.
 - [243] Y. Huang et al. Metatool benchmark for large language models: Deciding whether to use tools and which to use. *arXiv preprint arXiv:2310.03128*, 2023.
 - [244] P. Lu et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
 - [245] K. Wang et al. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.
 - [246] C. Rawles et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.
 - [247] Z. Xing et al. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.
 - [248] Z. Liu et al. Infi-mmr: Curriculum-based unlocking multimodal reasoning via phased reinforcement learning in multimodal small language models. *arXiv preprint arXiv:2505.23091*, 2025.
 - [249] X. Liu et al. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025.
 - [250] T. Xiao et al. Advancing multimodal reasoning capabilities of multimodal large language models via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025.
 - [251] H. Yao et al. Dense connector for mllms. *Advances in Neural Information Processing Systems*, 37:33108–33140, 2024.
 - [252] H. Jin et al. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024.
 - [253] P. Zhang et al. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.
 - [254] W. Hou et al. Memory-augmented multimodal llms for surgical vqa via self-contained inquiry. *arXiv preprint arXiv:2411.10937*, 2024.
 - [255] J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
 - [256] T. Kojima et al. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
 - [257] W. Wang et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.
 - [258] L. Du et al. Mm-prm: Enhancing multimodal mathematical reasoning with scalable step-level supervision. *arXiv preprint arXiv:2505.13427*, 2025.
 - [259] X. Chen et al. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*, 2025.
 - [260] Z. Kang et al. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025.
 - [261] R. Zhang et al. Accelerating best-of-n via speculative rejection. In *ICML 2024 Workshop on Structured Probabilistic Inference* {&} *Generative Modeling*, 2024.
 - [262] S. Yao et al. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.
 - [263] Z. Bi et al. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*, 2024.
 - [264] Y. Wang et al. Visuothink: Empowering lvlm reasoning with multimodal tree search, 2025.
 - [265] R. Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pp. 72–83. Springer, 2006.
 - [266] J. Wu et al. Boosting multimodal reasoning with automated structured thinking. *arXiv preprint arXiv:2502.02339*, 2025.
 - [267] R. Zhang et al. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024.
 - [268] J. Guo et al. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.
 - [269] Q. Chen et al. M³cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024.
 - [270] Z. Zhang et al. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.
 - [271] J. Jiang et al. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. *arXiv preprint arXiv:2507.07424*, 2025.
 - [272] B. Luan et al. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*, 2024.
 - [273] H. Shao et al. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.
 - [274] Y. Dong et al. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9062–9072, 2025.
 - [275] C. Mitra et al. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.
 - [276] O. Thawakar et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms, 2025.
 - [277] Y. Wang et al. R1-compress: Long chain-of-thought compression via chunk compression and search. *arXiv preprint arXiv:2505.16838*, 2025.
 - [278] H. Khalaf et al. Inference-time reward hacking in large language models. *arXiv preprint arXiv:2506.19248*, 2025.

- [279] C. Wang et al. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025.
- [280] Y. Liu et al. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.
- [281] Y. Peng et al. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.
- [282] Y. Zhan et al. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025.
- [283] Q. Yin et al. Tiny-r1v: Lightweight multimodal unified reasoning model via model merging, 2025.
- [284] W. Huang et al. Mapo: Mixed advantage policy optimization. *arXiv preprint arXiv:2509.18849*, 2025.
- [285] H. Deng et al. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.
- [286] N. Shinn et al. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- [287] H. Zhang et al. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- [288] Y. LeCun et al. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.
- [289] L. Xu et al. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.
- [290] C. Packer et al. Memgpt: Towards llms as operating systems. 2023.
- [291] E. Song et al. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024.
- [292] L. Fan et al. Posterior-grpo: Rewarding reasoning processes in code generation. *arXiv preprint arXiv:2508.05170*, 2025.
- [293] Y. Chen et al. R1-code-interpreter: Training llms to reason with code via supervised and reinforcement learning. *arXiv preprint arXiv:2505.21668*, 2025.
- [294] H. Wang et al. Otc: Optimal tool calls via reinforcement learning. *arXiv e-prints*, pp. arXiv–2504, 2025.
- [295] N. Shang et al. rstar2-agent: Agentic reasoning technical report. *arXiv preprint arXiv:2508.20722*, 2025.
- [296] R. Xu et al. Medagentgym: Training llm agents for code-based medical reasoning at scale. *arXiv preprint arXiv:2506.04405*, 2025.
- [297] Z. Liu et al. Ml-agent: Reinforcing llm agents for autonomous machine learning engineering. *arXiv preprint arXiv:2505.23723*, 2025.
- [298] J. Wu et al. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025.
- [299] C. Wang et al. Mllm-tool: A multimodal large language model for tool agent learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6678–6687. IEEE, 2025.
- [300] Z. Gao et al. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. *arXiv preprint arXiv:2412.15606*, 2024.
- [301] OpenAI. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>, 2025.
- [302] R. Nakano et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [303] Z. Zhang et al. Vision search assistant: Empower vision-language models as multimodal search engines. *arXiv preprint arXiv:2410.21220*, 2024.
- [304] Z. Hu et al. Avis: Autonomous visual information seeking with large language models. *arXiv preprint arXiv:2306.08129*, 3, 2023.
- [305] W. Chen et al. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022.
- [306] Z. Hu et al. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23369–23379, 2023.
- [307] Z. Chen et al. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*, 2024.
- [308] H. Wang et al. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. *arXiv preprint arXiv:2509.02544*, 2025.
- [309] M. Chen et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [310] Y. Li et al. Competition-level code generation with alpha-code. *Science*, 378(6624):1092–1097, 2022.
- [311] D. Guo et al. Deepseek-coder: When the large language model meets programming—the rise of code intelligence, 2024. URL <https://arxiv.org/abs/2401.14196>, 5:19, 2024.
- [312] H. Le et al. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.
- [313] P. Shojaei et al. Execution-based code generation using deep reinforcement learning. *arXiv preprint arXiv:2301.13816*, 2023.
- [314] Y. Chen et al. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*, 2025.
- [315] Y. Feng et al. Towards better correctness and efficiency in code generation. *arXiv preprint arXiv:2508.20124*, 2025.
- [316] S. Dou et al. StepCoder: Improve code generation with reinforcement learning from compiler feedback. *arXiv preprint arXiv:2402.01391*, 2024.
- [317] Y. Ye et al. Process-supervised reinforcement learning for code generation. *arXiv preprint arXiv:2502.01715*, 2025.
- [318] Y. Wang et al. Co-evolving llm coder and unit tester via reinforcement learning. *arXiv preprint arXiv:2506.03136*, 2025.
- [319] J. Wang et al. Repogenreflex: Enhancing repository-level code completion with verbal reinforcement and retrieval-augmented generation. *arXiv preprint arXiv:2409.13122*, 2024.
- [320] H. Lin et al. Os-r1: Agentic operating system kernel tuning with reinforcement learning. *arXiv preprint arXiv:2508.12551*, 2025.
- [321] OpenAI. Thinking with images. <https://openai.com/index/thinking-with-images/>, 2025.
- [322] H. Shao et al. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.
- [323] A. Kirillov et al. Segment anything, 2023.
- [324] S. Liu et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.
- [325] C. Duan et al. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*, 2025.
- [326] Z. Guo et al. Can we generate images with cot? let’s verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.
- [327] C. Tong et al. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint*

- arXiv:2505.17017*, 2025.
- [328] A. Das et al. Neural modular control for embodied question answering. In *Conference on robot learning*, pp. 53–62. PMLR, 2018.
- [329] D. S. Chaplot et al. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.
- [330] D. Jayaraman and K. Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1238–1247, 2018.
- [331] M. Shoenybi et al. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- [332] M. H. Daniel Han and U. team. Unsloth, 2023.
- [333] B. Chen et al. Fireact: Toward language agent fine-tuning, 2023.
- [334] A. Zeng et al. Agenttuning: Enabling generalized agent abilities for llms, 2023.
- [335] S. Diao et al. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. *arXiv preprint arXiv:2306.12420*, 2023.
- [336] L. von Werra et al. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020.
- [337] Hugging Face. Open rl: A fully open reproduction of deepseek-rl, January 2025.
- [338] J. Hu et al. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.
- [339] T. Xie et al. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning, 2025.
- [340] Y. Zheng et al. Easyrl: An efficient, scalable, multi-modality rl training framework. <https://github.com/hyoyuga/EasyRL>, 2025.
- [341] W. Zeng et al. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. <https://hkust-nlp.notion.site/simplerl-reason>, 2025. Notion Blog.
- [342] L. Wen et al. Light-rl: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*, 2025.
- [343] W. Fu et al. Areal: A large-scale asynchronous reinforcement learning system for language reasoning, 2025.
- [344] Z. Liu et al. Visual agentic reinforcement fine-tuning, 2025.
- [345] B. Jin et al. Search-rl: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.
- [346] Z. Wang et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning, 2025.
- [347] K. Zhang et al. Marti: A framework for multi-agent llm systems reinforced training and inference, 2025.
- [348] M. F. M. Team and M. A. I. Team. Mirorl: An mcp-first reinforcement learning framework for deep research agent. <https://github.com/MiroMindAI/MiroRL>, 2025.
- [349] C. Yu et al. Aworld: Orchestrating the training recipe for agentic ai, 2025.
- [350] G. Sheng et al. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.
- [351] Y. Yang et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.
- [352] S. Chen et al. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025.
- [353] X. Liu et al. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*, 2025.
- [354] H. Song et al. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025.
- [355] Y. Liu et al. rstar-coder: Scaling competitive code reasoning with a large-scale verified dataset. *arXiv preprint arXiv:2505.21297*, 2025.
- [356] X. Zhang et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025.
- [357] S. Yang et al. Instructvla: Vision-language-action instruction tuning from understanding to manipulation. *arXiv preprint arXiv:2507.17520*, 2025.
- [358] J. Wu et al. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025.
- [359] Y. Liu et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.
- [360] J. Roberts et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.
- [361] B. Y. Lin et al. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.
- [362] R. Qiao et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.
- [363] H. Rasheed et al. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025.
- [364] Z. Wang et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.
- [365] A. Maharana et al. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024.
- [366] H. Wu et al. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.
- [367] W. Wang et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.
- [368] L. Phan et al. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*, 2025.
- [369] Q. Wang et al. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*, 2025.
- [370] T. Ge et al. Advancing vision-language models in front-end development via data synthesis. *arXiv preprint arXiv:2503.01619*, 2025.
- [371] P. Wu and S. Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.
- [372] W. Wang et al. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.
- [373] K. Cheng et al. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024.
- [374] K. Li et al. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*, 2025.
- [375] W. Li et al. On the effects of data scale on ui control agents. *Advances in Neural Information Processing Systems*, 37:92130–92154, 2024.
- [376] R. Kapoor et al. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *European Conference on Computer Vision*, pp. 161–178. Springer, 2024.
- [377] Y. Dong et al. Ha-vln: A benchmark for human-aware

- navigation in discrete-continuous environments with dynamic multi-human interactions, real-world validation, and an open leaderboard. *arXiv preprint arXiv:2503.14229*, 2025.
- [378] R. Xu and J. Peng. A comprehensive survey of deep research: Systems, methodologies, and applications, 2025.
- [379] Y. Huang et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025.
- [380] W. Zhang et al. From web search towards agentic deep research: Incentivizing search with reasoning agents. *arXiv preprint arXiv:2506.18959*, 2025.
- [381] Y. Zheng et al. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments, 2025.
- [382] Microsoft. Introducing researcher and analyst in microsoft 365 copilot. <https://www.microsoft.com/en-us/microsoft-365/blog/2025/03/25/introducing-researcher-and-analyst-in-microsoft365-copilot/>, 2025.
- [383] Moonshot AI. Kimi-researcher: End-to-end rl training for emerging agentic capabilities. <https://moonshotai.github.io/Kimi-Researcher/>, 2025.
- [384] Zhipu AI. Autoglm rumination. <https://autoglm-research.zhipuai.cn/>, 2025.
- [385] Manus Team. Manus: General ai agent that bridges mind and action. <https://manus.im/app>, 2025.
- [386] Q. Bu et al. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.
- [387] Q. Lv et al. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*, 2025.
- [388] K. Black et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- [389] C. Cheang et al. Gr-3 technical report. *arXiv preprint arXiv:2507.15493*, 2025.
- [390] K. Black et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.
- [391] J. Lee et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.
- [392] J. Perlo et al. Embodied ai: Emerging risks and opportunities for policy action. *arXiv preprint arXiv:2509.00117*, 2025.
- [393] Y. Liu et al. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025.
- [394] C. Li et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.
- [395] H. Zhang et al. Huatuoqpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.
- [396] J. Chen et al. Huatuoqpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*, 2023.
- [397] J. Chen et al. Huatuoqpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.
- [398] J. Chen et al. Huatuoqpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.
- [399] P. Xia et al. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024.
- [400] A. Yu et al. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*, 2025.
- [401] W. Zhang et al. Patho-agenticrag: Towards multimodal agentic retrieval-augmented generation for pathology vlms via reinforcement learning. *arXiv preprint arXiv:2508.02258*, 2025.
- [402] X. Li et al. At-cxr: Uncertainty-aware agentic triage for chest x-rays. *arXiv preprint arXiv:2508.19322*, 2025.
- [403] X. Lan et al. Gem: Empowering mllm for grounded eeg understanding with time series and images. *arXiv preprint arXiv:2503.06073*, 2025.
- [404] K. Su et al. A fully autonomous robotic ultrasound system for thyroid scanning. *Nature Communications*, 15(1):4004, 2024.
- [405] A. Pore et al. Safe reinforcement learning using formal verification for tissue retraction in autonomous robotic-assisted surgery. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4025–4031. IEEE, 2021.
- [406] J. Liu et al. Robotic-assisted navigation system for pre-operative lung nodule localization: a pilot study. *Translational Lung Cancer Research*, 12(11):2283, 2023.
- [407] J. Ye et al. Mobile-agent-v3: Foundational agents for gui automation. *arXiv preprint arXiv:2508.15144*, 2025.
- [408] S. Nayak et al. Ui-vision: A desktop-centric gui benchmark for visual perception and interaction. *arXiv preprint arXiv:2503.15661*, 2025.
- [409] H. H. Zhao et al. Worldgui: An interactive benchmark for desktop gui automation from any starting point. *arXiv preprint arXiv:2502.08047*, 2025.
- [410] Q. Lu et al. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv preprint arXiv:2406.08451*, 2024.
- [411] Q. Yang et al. Fingertip 20k: A benchmark for proactive and personalized mobile llm agents. *arXiv preprint arXiv:2507.21071*, 2025.
- [412] Y. Cui et al. Chain-of-thought for autonomous driving: A comprehensive survey and future prospects. *arXiv preprint arXiv:2505.20223*, 2025.
- [413] Z. Yuan et al. Autodrive-r²: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*, 2025.
- [414] A. Ishaq et al. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. *arXiv preprint arXiv:2503.10621*, 2025.
- [415] Y. Liu et al. Rec-gpt4v: Multimodal recommendation with large vision-language models. *arXiv preprint arXiv:2402.08670*, 2024.
- [416] C. Huang et al. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734*, 2025.
- [417] H. Luo et al. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025.
- [418] H. Luo et al. Ada-r1: Hybrid-cot via bi-level adaptive reasoning optimization. *arXiv preprint arXiv:2504.21659*, 2025.
- [419] W. Xiao et al. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*, 2025.
- [420] C. Lou et al. Adacot: Pareto-optimal adaptive chain-of-thought triggering via reinforcement learning. *arXiv preprint arXiv:2505.11896*, 2025.
- [421] T. Gu et al. Mllmgaurd: A multi-dimensional safety evaluation suite for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:7256–7295, 2024.
- [422] Z. Ying et al. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*, 2024.
- [423] X. Liu et al. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024.
- [424] X. Yang et al. Mla-trust: Benchmarking trustworthiness of multimodal llm agents in gui environments. *arXiv*

preprint arXiv:2506.01616, 2025.

- [425] R. Pi et al. Mllm-protector: Ensuring mllm’s safety without hurting performance. *arXiv preprint arXiv:2401.02906*, 2024.
- [426] S. Raza et al. Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems. *arXiv preprint arXiv:2506.04133*, 2025.