Reinforcement Fine-Tuning Powers Reasoning Capability of Multimodal Large Language Models

Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifu Luo, Yifei Zhao, Kai Qin, Xufei Lv,
Tiantian Zhang, Yongzhe Chang, Xueqian Wang
Tsinghua Shenzhen International Graduate School, Tsinghua University
sun-hy23@mails.tsinghua.edu.cn
Project: https://github.com/Sun-Haoyuan23/Awesome-RL-based-Reasoning-MLLMs

Abstract

Standing in 2025, at a critical juncture in the pursuit of Artificial General Intelligence (AGI), reinforcement fine-tuning (RFT) has demonstrated significant potential in enhancing the reasoning capability of large language models (LLMs) and has led to the development of cutting-edge AI models such as OpenAI-o1 and DeepSeek-R1. Moreover, the efficient application of RFT to enhance the reasoning capability of multimodal large language models (MLLMs) has attracted widespread attention from the community. In this position paper, we argue that reinforcement fine-tuning powers the reasoning capability of multimodal large language models. To begin with, we provide a detailed introduction to the fundamental background knowledge that researchers interested in this field should be familiar with. Furthermore, we meticulously summarize the improvements of RFT in powering reasoning capability of MLLMs into five key points: diverse modalities, diverse tasks and domains, better training algorithms, abundant benchmarks and thriving engineering frameworks. Finally, we propose five promising directions for future research that the community might consider. We hope that this position paper will provide valuable insights to the community at this pivotal stage in the advancement toward AGI. Summary of works done on RFT for MLLMs is available at the project.

1 Introduction

"The senses are the organs by which man perceives the world, and the soul acts through them as through tools." —— *Leonardo da Vinci*

Reinforcement learning (RL), a series of machine learning approaches in which an agent learns optimal decision-making strategies by continuously employing the trial-and-error paradigm [1]. Over the past four decades, from classic algorithms to deep neural networks, from value-based to policy-based methods, this fascinating field has witnessed consistent and substantial advancements through dedicated research and exploration. Currently at the threshold of 2025, Proximal Policy Optimization (PPO) [2] stands out as one of the most influential RL algorithms within the community.

Since the 2020s, the emergence of large language models (LLMs) has rapidly accelerated advancements across numerous interdisciplinary fields. Their remarkable zero-shot capabilities, along with emerging reasoning and planning abilities, have offered a glimpse of the potential for achieving Artificial General Intelligence (AGI). The reinforcement learning from human feedback (RLHF) [3] pipeline has facilitated the development of epoch-making models, exemplified by GPT-4 [4] and LLaMA [5, 6]. However, their intellectual capabilities are largely constrained by human annotators. This limitation is particularly evident in the models' reasoning abilities, specifically in their systematic capacity to derive logical inferences, methodically solve complex problems, and

Preprint. Under review.

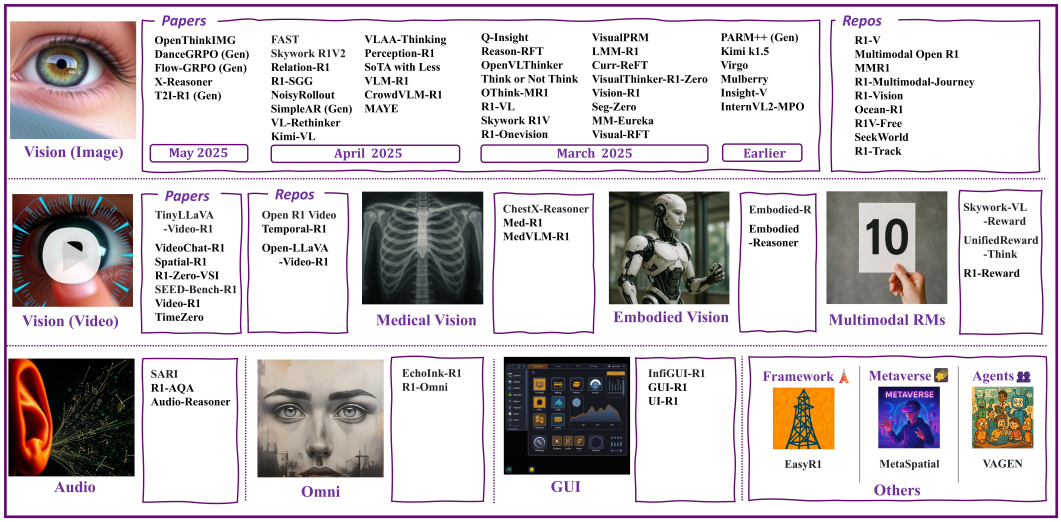

Figure 1: An overview of works done on reinforcement fine-tuning (RFT) for multimodal large language models (MLLMs). Works are sorted by release time and are collected up to May 15, 2025. Further detailed summary is provided in Appendix A.

effectively transfer knowledge across diverse domains. Subsequently, OpenAI-o1 [7] successfully applied large-scale reinforcement learning to the model training, which has significantly enhanced its reasoning capabilities. Moreover, DeepSeek-R1-Zero [8] has demonstrated remarkable self-evolution capabilities through a pure reinforcement learning process; additionally, DeepSeek-R1 [8] further stabilizes the RL process through cold start, ultimately exhibiting unparalleled reasoning capabilities. Their success has demonstrated the effectiveness of Reinforcement Fine-Tuning (RFT) in enhancing the reasoning capabilities of LLMs. However, it is noteworthy that their reasoning process involves only the textual modality.

Just as human perception depends on the harmonious interplay of multiple senses to form a coherent understanding of the world, multimodal large language models (MLLMs) integrate diverse data modalities (such as vision, text, audio, and so on) to perceive and reason about the complex and multimodal environments. Within this context, RFT has served as a critical mechanism for powering MLLMs with robust reasoning capabilities. With the emergence of DeepSeek-R1 [8], the efficient application of such training paradigm to enhance MLLM reasoning capabilities has attracted widespread attention from the community. A summary of these recent works is provided in Figure 1 and Appendix A. They have demonstrated that RFT significantly enhances the reasoning abilities of MLLMs, making them more proficient across diverse modalities, tasks, and domains. Our position is:

POSITION

Reinforcement Fine-Tuning (RFT) Powers Reasoning Capability of Multimodal Large Language Models (MLLMs).

Standing at the pivotal moment of 2025, we believe that the field of MLLM reasoning is experiencing an exciting period of transformation. Despite the challenges faced, this era presents unique opportunities for significant advancements. This position paper aims to provide the community with a valuable reference. Specifically, we are dedicated to answering the following three questions:

1. ***What background should researchers interested in this field know?*** We answer this question in Section 2, which can be divided into three parts. To enhance readers' understanding of this topic, we begin with the fundamental concepts and classic algorithms of reinforcement learning in Section 2.1, which are generally categorized into value-based and policy-based methods. Furthermore, following the recent survey [9], we present a concise overview of the current state of multimodal reasoning in Section 2.2, showing the trend from language-centric multimodal reasoning to collaborative multimodal reasoning. Finally, in Section 2.3, we provide a comprehensive discussion of representative RFT algorithms, emphasizing their similarities and differences; moreover, partly drawing on the recent survey [10], we categorize them as Critic-Model-Driven and Critic-Model-Free algorithms.

2. ***What has the community done?*** We meticulously answer this question from five perspectives, as outlined in Section 3. Firstly, the community has made significant process in the reasoning of diverse modalities. Secondly, significant progress has also been achieved across diverse tasks and domains. Thirdly, the community has advanced the development of better training algorithms. Fourthly, we reveal several exciting trends in the development of abundant reasoning multimodal benchmarks. Finally, thriving engineering frameworks have been continuously developed by the community.

3. ***What could the community do next?*** We critically answer this question by presenting five key points, as demonstrated in Section 4. Firstly, the community could further explore strategies for enhancing generalization across various modalities, tasks, and domains. Secondly, effectively combining the strengths of the outcome reward paradigm and the process reward paradigm is a promising direction for further research. Thirdly, we hope the community could devote increased attention to the safety of reasoning MLLMs. Fourthly, given the scarcity of multimodal data, further research into data augmentation techniques for reasoning scenarios is highly promising. Finally, the community could continue to conduct in-depth research on more effective algorithms, improved reward paradigms, and broader applications.

2 Background

2.1 Reinforcement Learning: Value-based and Policy-based Methods

Markov Decision Process. The Markov Decision Process (MDP) [11, 1] is a foundational mathematical framework for modeling sequential decision-making in stochastic environments. It extends Markov chains by incorporating actions and rewards, enabling agents to learn optimal policies that maximize cumulative long-term rewards. Formally, an MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \rho, \gamma)$. Herein, $\mathcal{S}$ represents the state space; $\mathcal{A}$ donates the action space; and $\mathcal{P}$ is the state transition function. At any time step t, for the state $s_t, s_{t+1} \in \mathcal{S}$ and the action $a_t \in \mathcal{A}$, $\mathcal{P}(s_{t+1}|s_t, a_t)$ denotes the probability of reaching state s_{t+1} after performing the action a_t in state s_t. $\mathcal{R}$ is the reward function, where $\mathcal{R}(s_t, a_t) \in \mathbb{R}$. ρ denotes the initial state distribution; and γ serves as the discount factor, indicating the importance of future rewards in relation to the current state. The agent-environment interaction typically operates in a blocking paradigm. At step t, the agent observes state s_t and, based on its policy $\pi(\cdot|s_t)$, performs action a_t. After taking the action, the environment updates to a new state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$, and yielding a reward $\mathcal{R}(s_t, a_t)$ for the agent. Reinforcement learning aims to acquire an optimal policy, denoted as $\pi^*(\cdot|s_t)$, by maximizing the cumulative discounted reward (also known as the return), $G_t = \sum_{k=t}^{T} \gamma^{k-t} \mathcal{R}(s_k, a_k)$, where T represents the episode horizon.

Classic Reinforcement Learning. Since the 1980s, prominent computer scientists Andrew G. Barto and Richard S. Sutton have laid theoretical foundations and developed key algorithms in the field of reinforcement learning. Recently, the 2024 ACM A.M. Turing Award was conferred upon them in recognition of their fundamental contributions to the field of reinforcement learning. Furthermore, their book "*Reinforcement Learning: An Introduction*" [1] has earned the title of "Bible of Reinforcement Learning" among researchers. Hence, we argue that a fundamental understanding of classic reinforcement learning is crucial for practitioners employing reinforcement fine-tuning techniques, especially to avoid ambiguity in conceptual definitions. We restate the definitions:

Definition 2.1 (State Value Function [1])**.** ***The value of a state*** s ***under a policy*** π*, denoted* $V_\pi(s)$*, is the expected return when starting in* s *and following* π *thereafter. For MDPs,* $V_\pi(s)$ *can be defined as:*

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}_{t+k+1} \,\middle|\, S_t = s\right]. \tag{1}$$

Definition 2.2 (Action Value Function [1])**.** ***The value of taking action*** a ***in state*** s ***under policy*** π*, denoted* $Q_\pi(s,a)$*, as expected return starting from* s*, taking action* a*, thereafter following policy* π*:*

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]. \tag{2}$$

Classic reinforcement learning algorithms are categorized as Value-based methods and Policy-based methods. Value-based methods primarily focus on learning a value function (usually the action value function $Q_\pi(s,a)$), from which a policy is subsequently derived. Early typical algorithms within such

paradigm include Q-Learning [1, 12] and SARSA [1]. With the rise of deep learning, researchers have increasingly employed neural networks to approximate the value function, leading to a series of improvements, exemplified by Deep Q-Network (DQN) [13, 14], Double DQN [15], Dueling DQN [16], Rainbow [17], and so on. In contrast, Policy-based methods directly and explicitly learn a target policy with the primary objective of identifying an optimal policy that maximizes the expected reward within the environment. REINFORCE [18] pioneers this paradigm, defining the target function as $J(\theta) = \mathbb{E}_{s_0}[V_{\pi_\theta}(s_0)]$, where s_0 is the inital state; subsequently, the objective function is differentiated with respect to the policy parameter θ, and gradient ascent method is applied to maximize this function. Actor-Critic algorithms [1, 19, 20] take a further step by fitting a value function to guide policy learning. Critic (value module), learns to discriminate between effective and ineffective actions based on data sampled by Actor (policy module), subsequently guiding Actor's policy update; concurrently, as Actor trains, the distribution of environment interaction data shifts, necessitating rapid adaptation for Critic to provide accurate value estimations under the evolving data distribution. It's important to clarify that Actor-Critic family algorithms are essentially Policy-based algorithms: as they fundamentally aim to optimize the policy while only concurrently learning a value function to enhance policy learning efficiency.

From TRPO to PPO. Trust Region Policy Optimization (TRPO) [21] and Proximal Policy Optimization (PPO) [2] inherit the Actor-Critic paradigm; therefore, they are actually Policy-based methods. A significant limitation of previous Policy-based methods is the potential for abrupt policy degradation when updating parameters along the policy gradient (often attributed to excessively large step sizes). TRPO promotes stable and effective policy learning by employing a "trust region" constraint during policy updates; and it theoretically guarantees monotonic improvement in policy learning. We present the theoretical policy optimization objective of TRPO in Proposition 2.1 without any further proof.

Proposition 2.1 (TRPO Objective [21])**.** *The current policy, parameterized by θ_{old}, is denoted as $\pi_{\theta_{old}}$. Primary goal is to find a better policy π_θ utilizing current policy $\pi_{\theta_{old}}$. The objective is:*

$$\max_\theta \mathbb{E}_{s\sim\upsilon^{\pi_{\theta_{old}}}} \mathbb{E}_{a\sim\pi_{\theta_{old}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a) \right] \quad s.t.\ \mathbb{E}_{s\sim\upsilon^{\pi_{\theta_{old}}}} \left[\mathbb{D}_{KL}(\pi_{\theta_{old}}(\cdot|s), \pi_\theta(\cdot|s)) \right] \leq \delta, \tag{3}$$

where υ^π is the state visitation distribution under policy π; $A_\pi(s,a)$ is the advantage function with definition of $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$; $\mathbb{D}_{KL}$ is the Kullback-Leibler divergence that serves as a constraint to ensure that the new policy remains sufficiently close to the old policy.

However, TRPO practically solves the objective using methods such as Taylor expansion approximation, conjugate gradient, and line search, making computational cost of each update step significantly high. Thereby, the Proximal Policy Optimization (PPO) [2] is proposed, directly incorporating first-order optimization through a clipped objective function (PPO-Clip) or a KL divergence penalty term (PPO-Penalty), avoiding second-order calculation. PPO objectives are presented in Proposition 2.2.

Proposition 2.2 (PPO Objectives [2])**.** *PPO-Penalty incorporates KL divergence constraint into the objective function using Lagrangian multiplier method; and coefficient β is updated during training:*

$$\max_\theta \mathbb{E}_{s\sim\upsilon^{\pi_{\theta_{old}}}} \mathbb{E}_{a\sim\pi_{\theta_{old}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a) - \beta \mathbb{D}_{\mathrm{KL}}(\pi_{\theta_{old}}(\cdot|s), \pi_\theta(\cdot|s)) \right] \tag{4}$$

Setting $d = \mathbb{E}_{s\sim\upsilon^{\pi_{\theta_{old}}}}[\mathbb{D}_{KL}(\pi_{\theta_{old}}(\cdot|s), \pi_\theta(\cdot|s))]$. If $d < d_{targ}/1.5$, $\beta \leftarrow \beta/2$; if $d > d_{targ} \times 1.5$, $\beta \leftarrow 2\beta$. PPO-Clip restricts the objective function more directly with the clip operation:

$$\max_\theta \mathbb{E}_{s\sim\upsilon^{\pi_{\theta_{old}}}} \mathbb{E}_{a\sim\pi_{\theta_{old}}} \left[\min \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a), \mathrm{clip} \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}, 1-\varepsilon, 1+\varepsilon \right) A_{\pi_{\theta_{old}}}(s,a) \right) \right] \tag{5}$$

For PPO-Clip, as shown in Equation (5), if the advantage function is positive (action value is higher than average), maximizing the equation leads to an increase of $\pi_\theta/\pi_{\theta_{\text{old}}}$, but it is constrained not to exceed $1+\varepsilon$; similarly, if the advantage function is negative (action value is lower than average), maximizing the equation leads to a decrease of $\pi_\theta/\pi_{\theta_{\text{old}}}$, but it is constrained not to exceed $1-\varepsilon$.

2.2 Multimodal Reasoning: Language-Centric and Collaborative Paradigms

Recent advances in the reasoning capabilities of Large Language Models (LLMs) [22–28] pave a significant step towards achieving Artificial General Intelligence (AGI), exemplified by OpenAI-o1

[7], DeepSeek-R1 [8], and so on. Furthermore, sustained success of Multimodal Large Language Models (MLLMs) has motivated researchers to explore the integration of LLM reasoning capabilities with multimodal processing [29, 30, 9, 10, 31], resulting in notable implementations such as Kimi k1.5 [32], OpenAI o3 and o4-mini [33], Grok 3 [34], and so on. In the recent survey [9], the authors proposed a taxonomy of multimodal reasoning into two categories: the Language-Centric Multimodal Reasoning and the Collaborative Multimodal Reasoning. We argue that such taxonomy is insightful and constructive; furthermore, such taxonomy effectively illustrates an evolution in multimodal reasoning technologies, specifically a shift from language-dominated control to collaborative multimodal co-reasoning. Hence, we further follow this taxonomy [9]:

Language-Centric Multimodal Reasoning [9]. In this paradigm, the multimodality (beyond language) primarily functions to acquire perceptual information and extract features; the reasoning process, on the contrary, is predominantly driven by the language modality. This paradigm is further divided into the One-pass Multimodality Perception and the Active Multimodality Perception based on the multimodal perception triggering mechanisms. One-pass Multimodality Perception methods treat multimodal information (beyond language) as static context, encoding them only once during the model's input stage. In contrast, for Active Multimodality Perception methods, language modality's generation of intermediate reasoning steps triggers iterative multimodal re-perception cycles.

Collaborative Multimodal Reasoning [9]. In this paradigm, the reasoning process further necessitates multimodal (beyond language) action reasoning and multimodal (beyond language) state updating, and multimodal representation extends beyond passive perception to active collaboration with the language modality throughout the reasoning process. Regarding the multimodal (beyond language) action reasoning, it transcends purely linguistic instructions and generates internal reasoning actions autonomously; furthermore, it exhibits explicit reasoning trajectories within the multimodal feature space. Then, the model dynamically updates multimodal contextual information by executing the aforementioned multimodal (beyond language) actions, which actually introduce new constraints to the language modality and thereby trigger subsequent multimodal reasoning steps.

2.3 Reinforcement Fine-Tuning: Critic-Model-Driven and Critic-Model-Free Algorithms

Leveraging reinforcement fine-tuning (RFT), recent studies have introduced novel post-training algorithms to enhance reasoning capabilities of LLMs [28, 7, 8] and MLLMs [32–34]. In the recent survey [10], the authors categorized RL-based training into value-based and value-free algorithms. While this specific point is correct within the RFT context, we argue that it could conflict with the concepts of classic reinforcement learning (for example, PPO is classified as a value-based algorithm within this system; however, it is actually a Policy-based algorithm). Herein, we refine the taxonomy for reinforcement fine-tuning as: Critic-Model-Driven algorithms and Critic-Model-Free algorithms.

Critic-Model-Driven algorithms. Since the introduction of PPO [2] in 2017, it has become one of the most popular actor-critic RL algorithms for policy optimization of LLMs [3, 35] and MLLMs. Within the context of MLLMs, the input (m,t) consisting of both multimodal (beyond language) contents m and a textual query t; then, PPO objective in Equation (5) can be transferred as [28, 35]:

$$\max_{\theta} \quad \mathbb{E}_{((m,t),a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|(m,t))}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\right.\right.$$
$$\left.\left.\min\left(\frac{\pi_\theta(o_{i,t}|(m,t),o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|(m,t),o_{i,<t})}\hat{A}_{i,t}(\phi),\text{clip}\left(\frac{\pi_\theta(o_{i,t}|(m,t),o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|(m,t),o_{i,<t})},1-\varepsilon,1+\varepsilon\right)\hat{A}_{i,t}(\phi)\right)\right)\right], \tag{6}$$

where $\hat{A}_{i,t}(\phi)$ represents the Generalized Advantage Estimation (GAE) [36]: it is computed using the "value" provided by the critic model $V_\phi(o_{i,t}|(m,t),o_{i,<t})$ that is trained concurrently with the policy model $\pi_\theta(o|(m,t))$. In the domain of LLM policy optimization, several studies have made significant improvements along this line, exemplified by Open-Reasoner-Zero [35] and VC-PPO [37].

Critic-Model-Free algorithms. Group Relative Policy Optimization (GRPO) [8, 38] discards the critic model and the calculation of GAE in PPO by sampling and normalizing rewards within a group of G outputs, which significantly enhances efficiency and reduces memory consumption. In addition, a KL-divergence penalty is applied to constrain the optimized model $\pi_\theta(o|(m,t))$, mitigating excessive divergence from initial SFT model $\pi_{\text{ref}}(o|(m,t))$. We detail the GRPO objective in Proposition 2.3.

Proposition 2.3 (GRPO Objective [8, 38]). *For each sample $((m,t),a)$, GRPO samples a group of outputs $\{o_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{old}}(o|(m,t))$, and the policy model $\pi_\theta(o|(m,t))$ is optimized by:*

$$\max_{\theta} \mathbb{E}_{((m,t),a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|(m,t))} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(\frac{\pi_\theta(o_{i,t}|(m,t),o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|(m,t),o_{i,<t})}\hat{A}_{i,t}, \right.\right.\right. \\ \left.\left.\left. \text{clip}\left(\frac{\pi_\theta(o_{i,t}|(m,t),o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|(m,t),o_{i,<t})}, 1-\varepsilon, 1+\varepsilon \right)\hat{A}_{i,t} \right) - \beta\mathbb{D}_{\text{KL}}\left(\pi_\theta,\pi_{\text{ref}}\right)_{i,t} \right) \right], \tag{7}$$

where $\hat{A}_{i,t}=\widetilde{r}_i=\left(r(o_i,a)-\text{mean}\left(\left\{r(o_i,a)\right\}_{i=1}^{G}\right)\right)/\text{std}\left(\left\{r(o_i,a)\right\}_{i=1}^{G}\right)$ is the group relative reward (advantage); and $\left\{r(o_i,a)\right\}_{i=1}^{G}$ represents the rewards of response group $\{o_i\}_{i=1}^{G}$ that computed by reward function. Furthermore, the KL divergence in GRPO is calculated by the K3 estimator [39].

For LLM policy optimization, researchers have further proposed several representative improvements towards GRPO, exemplified by DAPO [40], Dr.GRPO [41], and so on. Furthermore, considerable engineering efforts within the LLM community have focused on adapting algorithms to achieve more stable and efficient training, exemplified by Light-R1 [42], TinyZero [43], and so on.

3 RFT for MLLMs: What has the community done?

Large Reasoning Models (LRMs) represent cutting-edge AI models designed to devote more time to thinking before providing a response, thus achieving superior reasoning capabilities. With the introduction of OpenAI-o1 [7], reinforcement fine-tuning has shown great potential in the domain of LLMs. However, the process reward paradigm it utilized exhibits unstable training and limited generalizability [44]. Furthermore, applying this paradigm to MLLMs might be further hindered by the challenge of generalization across diverse tasks. VisualPRM [45] and PARM++ [46] represent pioneering efforts in this area, demonstrating its great potential for future developments. Inspired by the success of DeepSeek-R1 [8], the community suggests that simple rule-based rewards (i.e. the outcome reward paradigm), even without a separate learned reward model, can suffice for the autonomous development of complex reasoning capabilities in LLMs [26–28]. Effectiveness of this paradigm is also rapidly and thoroughly validated within the MLLM community. In particular, since March 2025, substantial progress has been made in enhancing multimodal reasoning within this paradigm. Generally, reinforcement fine-tuning has achieved significant successes in powering the reasoning ability of MLLMs. We meticulously divide the successes into the following five points:

Success 1: Diverse Modalities. As demonstrated in Figure 1, recent advancements in reinforcement fine-tuning (RFT) for multimodal large language models (MLLMs) are summarized. It is indicated that RFT has significantly enhanced the reasoning abilities of vision, audio, omni-multimodal, graphical user interface (GUI), metaverse interaction and agents in MLLMs. Notably, in addition to substantial progress in the vision modality, the community has also achieved significant breakthroughs in other modalities. Audio-Reasoner [47], R1-AQA [48] and SARI [49] have utilized RFT to enhance the reasoning capabilities of Large Audio Language Models (LALMs) in Audio Question Answering (AQA) tasks. R1-Omni [50] and EchoInk-R1 [51] have successfully implemented RFT in Omni-multimodal large language models, which fundamentally rely on both visual and auditory modalities. UI-R1 [52], GUI-R1 [53] and InfiGUI-R1 [54] have similarly applied RFT to advance action prediction tasks for graphic user interface (GUI) agents, thereby enhancing their understanding and control capabilities. MetaSpatial [55] has made significant progress in employing RFT to enhance 3D spatial reasoning within metaverse scenarios. VAGEN [56] has advanced the training of VLM-based visual agents through a multi-turn RFT framework.

Success 2: Diverse Tasks and Domains (Take Vision Modality as an Example). As previously mentioned, RFT has achieved significant successes in diverse modalities. Furthermore, within the vision modality alone, considerable successes have been achieved across a wide range of tasks and domains. Mathematical visual reasoning [57–63] and academic multi-discipline reasoning [64–66], tasks that receive a lot of attention from the community, require the precise integration of symbolic processing, visual analysis, and logical reasoning. Much groundbreaking works on this subject have already been carried out by the community: InternVL2-MPO [67], Mulberry [68], Virgo [69], MM-EUREKA [70], Vision-R1 [71], LMM-R1 [72], VisualPRM [45], MMR1

[73], R1-Onevision [74], SkyworkR1V [75], R1-VL [76], OpenVLThinker [77], VL-Rethinker [78], NoisyRollout [79], Skywork R1V2 [80] and FAST [81]. Meanwhile, vision-driven tasks have also attracted widespread attention from the community: VLM-R1 [82] has presented evidence of the feasibility and effectiveness of applying RFT to visual understanding tasks like referring expression compression and open-vocabulary object detection; CrowdVLM-R1 [83] has adapted RFT to the task of crowd counting; VisualThinker-R1-Zero [84] has employed RFT to vision-centric spatial reasoning tasks; CLS-RL and No-Thinking-RL [85] have utilized RFT for the few-shot image classification task; Seg-Zero [86] has applied RFT to the image segmentation task; Q-Insight [87] has adapted RFT to the image quality assessment task; Perception-R1 [88] has employed RFT to the visual perception task; R1-SGG [89] and Relation-R1 [90] have utilized RFT for the tasks of scene graph generation and relation comprehension; R1-Track [91] has applied RFT to the visual object tracking task; SeekWorld [92] has adapted RFT to the visual geolocation reasoning task; V-ToolRL [93] has employed RFT for the adaptive invocation of external vision tools. Furthermore, a significant number of works have focused on multi-task and multi-domain joint training to simultaneously improve model performance across multiple tasks and domains: Insight-V [94], Visual-RFT [95], Reason-RFT [96], ThinkLite-VL [97], VLAA-Thinking [98], Kimi-VL-Thinking [99], R1-Vision [100] and Ocean-R1 [101]. In the specific domain of temporal vision (video), applications of RFT have successfully enhanced video reasoning capabilities: Open-R1-Video [102], TimeZero [103], Temporal-R1 [104], Open-LLaVA-Video-R1 [105], Video-R1 [106], SEED-Bench-R1 [107], R1-Zero-VSI [108], Spatial-R1 [109], VideoChat-R1 [110] and TinyLLaVA-Video-R1 [111]. Moreover, in specific domain disciplines, the application of RFT has also successfully enhanced the reasoning ability of domain-specific MLLMs: MedVLM-R1 [112], Med-R1 [113] and ChestX-Reasoner [114] have represented significant advancements in medical reasoning capabilities for MLLMs in medical vision; Embodied-Reasoner [115] and Embodied-R [116] have demonstrated substantial reasoning capabilities progress in embodied vision. Beyond that, RFT has also further powered multimodal generation (especially the text-to-image generation), exemplified by PARM++ [46], SimpleAR [117], T2I-R1 [118], Flow-GRPO [119] and DanceGRPO [120].

Success 3: Better Training Algorithms. In addition to exploring the application of GRPO across diverse modalities, tasks, and domains, we observe that the community has also conducted in-depth exploration into better algorithms. These explorations primarily focus on training paradigm [121, 70, 77–79], algorithmic strategy [122, 76, 81], and data selection [97]. Curr-ReFT [121] proposes a novel post-training paradigm comprising two stages: the curriculum reinforcement learning (difficulty-aware reward design) and the rejected sampling-based self-improvement (selective learning from high-quality examples). MM-EUREKA [70] introduces online filtering paradigm, which eliminates prompts that yield responses deemed either entirely correct or entirely incorrect during training; furthermore, the study's implementation of DAPO [40] and ADORA [123] also provides valuable insights for future improved training paradigms. OpenVLThinker [77] iteratively employs SFT and GRPO, utilizing reasoning data from previous iterations to achieve self-improvement; significantly, it evolves the training data to progressively include more challenging questions over iterations. VL-Rethinker [78] introduces the Selective Sample Replay (SSR) to mitigate the vanishing advantage issue in GRPO and incorporates the Forced Rethinking to explicitly enforce a self-reflection reasoning step. NoisyRollout [79] integrates trajectories from both clean and moderately distorted images to foster targeted diversity in visual perception and the resulting reasoning patterns; additionally, it employs a noise annealing schedule that progressively reduces distortion strength over training, maximizing the advantages of noisy signals in earlier phases while ensuring stability and scalability in later stages. OThink-MR1 [122] introduces GRPO-D, which enhances GRPO through the incorporation of a dynamic KL divergence strategy inspired by the ε-greedy strategy in classic reinforcement learning. R1-VL [76] introduces StepGRPO, which incorporates both the step-wise reasoning accuracy reward and the step-wise reasoning validity reward, thereby effectively mitigating the sparse reward challenge without applying process reward models. FAST [81] introduces FAST-GRPO that integrating three key components: model-based metrics for question characterization, an adaptive thinking reward mechanism, and difficulty-aware KL regularization. ThinkLite-VL [97] introduces an effective MCTS-based data filtering method that quantifies sample difficulty according to the number of iterations the model requires to solve each problem, thereby achieving state-of-the-art reasoning performance with fewer training samples.

Success 4: Abundant Benchmarks. As stated in the blog [124], abundant benchmarks have been essential on the path towards Artificial General Intelligence (AGI) in the future. In the domain of MLLM reasoning, especially in visual reasoning, there have long been general and recognized

benchmarks within the community. As recent surveys [29, 30, 9, 10, 31] have summarized them extensively, a detailed discussion of them is omitted here. Furthermore, our analysis reveals that following the emergence of DeepSeek-R1 [8], multimodal reasoning benchmarks have shown the following exciting six trends. The first trend is the increasing difficulty of benchmarks: for example, on the ZeroBench [125], all contemporary frontier MLLMs have completely failed in this regard. The second trend involves benchmarks that assess human-like reasoning capabilities: for example, V1-33K [126] evaluates the reasoning capabilities of MLLMs by implementing auxiliary tasks, a method frequently employed in human reasoning processes; GeoSense [127] evaluates the identification and adaptive application of geometric principles, which is an important human-like geometric reasoning mechanism that has been neglected in previous benchmarks; MM-IQ [128] evaluates the abstraction and reasoning abilities of MLLMs by utilizing human-like IQ tests. The third trend is more comprehensive benchmarks on classic domains: for example, MDK12-Bench [129] extends the data size and domain coverage of the multi-discipline domain; MV-MATH [130] expands the scope of mathematical reasoning, moving from single-visual contexts to encompass multi-visual scenarios; Spatial457 [131] innovatively broadens the scope of visual-spatial reasoning into six dimensions (6D); VCR-Bench [132] introduces the evaluation of video chain-of-thought (CoT) reasoning for video benchmarking; MME-CoT [133] further assesses the reasoning quality, robustness, and efficiency at a fine-grained level. The fourth trend is benchmarks for more realistic application scenarios: for example, Video-MMLU [134] assesses MLLMs on multi-discipline lecture tasks; GDI-Bench [135] evaluates MLLMs on document-specific reasoning tasks. The fifth trend is a transition from language-centric benchmarks to multimodal-centric (especially visual-centric) ones: for example, VisuLogic [136] represents a formidable visual reasoning benchmark that inherently poses significant difficulty to articulate in language. The sixth trend is the introduction of interactive elements: for example, iVISPAR [137] introduces a novel interactive benchmark designed to evaluate the spatial reasoning capabilities of VLMs acting as agents.

Success 5: Thriving Engineering Frameworks. Within the community, enhancements to engineering training frameworks have been pivotal in reducing research barriers and increasing development efficiency. Since the emergence of DeepSeek-R1 [8], several frameworks have significantly advanced the community's development. Open-R1-Multimodal has been a pioneering effort in this area that is built upon Open-R1 [138] and TRL [139], effectively implementing multimodal model training through the GRPO algorithm. R1-V [140] has taken a step further, making it support the Qwen2.5-VL model, the GEOQA task and the vLLM [141] for training acceleration. EasyR1 [142] is a clean fork of the original veRL [143] project. It features extensive support for models, algorithms, and datasets, along with support for padding-free training, checkpoint resumption, and tool integration. MAYA [144] offers a transparent and reproducible framework, along with a comprehensive evaluation scheme, for the application of RL to MLLMs; furthermore, it also serves as a lightweight and educational framework that elucidates the core logic of RL training.

4 Future Work: What could the community do next?

As discussed in Section 3, the emergence of Deepseek-R1 [8] has significantly increased interest within the community regarding utilization of RFT to further enhance the reasoning capabilities of MLLMs. Excitingly, the community has already achieved remarkable successes on this topic, including diverse modalities, diverse tasks and domains, better training algorithms, abundant benchmarks and thriving engineering frameworks. Following discussions with the MLLM, LLM, and RL communities, we believe that the following five points still warrant further research:

TO DO 1: Achieve Better Generalization across Modalities, Tasks and Domains. Although considerable research has focused on cross-task reasoning, existing efforts remain limited to specific domains and modalities; furthermore, the scope of these tasks is limited, typically encompassing only two or three tasks. However, in the pursuit of AGI, we have always aspired to develop a single model capable of adapting to a diverse array of modalities, tasks, and domains. Therefore, research on generalizable reasoning holds significant value. X-Reasoner [145] is a pioneer in this area, demonstrating that general-domain text-based post-training can enable generalizable reasoning and that performance in specialized domains could be further enhanced through training on domain-specific (e.g., medical-specific) text-only data. Moreover, it can be observed that there are still more points worth exploring in this area. Firstly, modalities other than textual and visual have not been addressed; therefore, future work could further explore generalizable reasoning capabilities

for more complex modalities, such as auditory and omni-multimodal. Furthermore, the reasoning capability generalization for broader tasks, such as progressing from the perceptual vision task (image) to the temporal vision task (video), deserves further exploration within the community. Lastly, generalization of the reasoning capability across broader domains, exemplified by the shift from general-domain to embodied-specific settings, remains an underexplored area that requires further systematic investigation.

TO DO 2: Combine the Outcome Reward Paradigm and the Process Reward Paradigm. The outcome reward paradigm offers high efficiency and ease of implementation; however, the sparsity of its rewards provides no intermediate feedback during the reasoning process. For the process reward paradigm, while dense rewards are available for intermediate reasoning steps, training of the process reward model (PRM) remains relatively challenging and unstable. Therefore, the community could consider integrating the outcome reward paradigm with the process reward paradigm. On one hand, PRM training could be powered by the outcome reward paradigm. Regarding multimodal reward model training, R1-Reward [146], UnifiedReward-Think [147] and Skywork-VL Reward [148] have conducted pioneering research, demonstrating that RFT could lead to more stable training dynamics and enhanced performance; therefore, future research could investigate the integration of outcome reward paradigm to enhance PRM training. On the other hand, further exploration is warranted regarding the provision of effective and dense rewards in the outcome reward paradigm. StepGRPO [76] represents a pioneering approach to this area, notably by incorporating dense step-wise rewards; however, it is limited to the vision mathematical reasoning task, and the applicability of such a methodology to other tasks, domains, and modalities requires further investigation.

TO DO 3: Pay More Attention towards the Safety of Reasoning MLLMs. Safeguarding MLLMs against security vulnerabilities and adversarial threats is a critical research area that has been widely explored in the community [149–152]. Recently, it has been indicated that reasoning LLMs present novel safety challenges stemming from their training algorithms, exposure to adversarial attacks during inference, and vulnerabilities inherent in their deployment environments [153–155]. Nevertheless, research specifically on safety for reasoning MLLMs remains notably limited, which is a critical area that demands increased attention from the community. Future research could further focus on developing advanced detection and defense mechanisms specifically designed for reasoning MLLMs. According to [28], this point can generally be divided into three components. Firstly, reward hacking, a persistent challenge within the community [156, 157], warrants further attention and effort. Moreover, the exploration of jailbreak attacks and defenses in reasoning MLLMs deserves greater focus within the community. Lastly, the issue of overthinking, as highlighted by pioneering works such as No-Thinking-RL [85] and FAST [81], is also a critical challenge within the community that could be further investigated across more diverse modalities, tasks, and domains.

TO DO 4: Investigate More Data Augmentation Attempts for Multimodality. Data augmentation (DA) has been demonstrated to be an effective technique for MLLMs' training [158, 159] and could potentially enhance the performance and robustness of models. In RFT settings for MLLMs, data is often scarce; therefore, internal data augmentation is likely to enhance the model's perception capabilities. NoisyRollout [79] pioneers in this area, which demonstrates that incorporating Gaussian noise during training could enhance the reasoning performance on visual mathematical task. Therefore, further exploration in the following points might be valuable. Firstly, appropriate DA methods for a broader range of visual tasks (such as the visual counting task) could be further explored. Moreover, it is also worthwhile to further explore more appropriate and diverse DA methods (such as RandomResizedCrop, RandomCrop, CenterCrop, RandFlip, RandomAffine, RandomInvert, and so on [159]) for all these tasks. Lastly, the potential for applying DA methods to other modalities and evaluating their effectiveness in these contexts merits further investigation.

TO DO 5: Explore Better Algorithms, Reward Paradigms, and Beyond. As previously discussed, the community has made substantial progress in developing improved training algorithms. Furthermore, this should continue to be one of the key areas of focus for community efforts. Regarding reward paradigms, rule-based rewards are typically employed in current algorithms. In future research, it is valuable to further explore automatic frameworks for designing task-specific reward functions. Finally, exploring the implementation of reinforcement fine-tuned reasoning MLLMs across diverse academic disciplines (such as architecture, aerospace, electric engineering, and so on) is a promising field that necessitates collaborative efforts from various disciplinary communities.

Acknowledgments and Disclosure of Funding

Haoyuan Sun extends his sincere gratitude to all community members who provided valuable supplementary support to the project. This work was supported by the Natural Science Foundation of Shenzhen (No.JCYJ20230807111604008, No.JCYJ20240813112007010), the Natural Science Foundation of Guangdong Province (No.2024A1515010003), National Key Research and Development Program of China (No.2022YFB4701400) and Cross-disciplinary Fund for Research and Innovation (No.JC2024002) of Tsinghua SIGS.

References

[1] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

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

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

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[5] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[6] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[7] OpenAI. Introducing openai o1. https://openai.com/o1/, 2024.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[9] Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. Mind with eyes: from language reasoning to multimodal reasoning. *arXiv preprint arXiv:2503.18071*, 2025.

[10] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025.

[11] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

[12] Francisco S Melo. Convergence of q-learning: A simple proof. *Institute Of Systems and Robotics, Tech. Rep*, pages 1–4, 2001.

[13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[15] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[16] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

[17] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[18] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[19] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[20] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016.

[21] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[22] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024.

[23] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

[24] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025.

[25] Jun Wang. A tutorial on llm reasoning: Relevant methods behind chatgpt o1. *arXiv preprint arXiv:2502.10867*, 2025.

[26] Dibyanayan Bandyopadhyay, Soham Bhattacharjee, and Asif Ekbal. Thinking machines: A survey of llm based reasoning strategies. *arXiv preprint arXiv:2503.10814*, 2025.

[27] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

[28] Chong Zhang, Yue Deng, Xiang Lin, Bin Wang, Dianwen Ng, Hai Ye, Xingxuan Li, Yao Xiao, Zhanfeng Mo, Qi Zhang, et al. 100 days after deepseek-r1: A survey on replication studies and more directions for reasoning language models. *arXiv preprint arXiv:2505.00551*, 2025.

[29] Yujie Lin, Ante Wang, Moye Chen, Jingyao Liu, Hao Liu, Jinsong Su, and Xinyan Xiao. Investigating inference-time scaling for chain of multi-modal thought: A preliminary study. *arXiv preprint arXiv:2502.11514*, 2025.

[30] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

[31] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang, Zhuotao Tian, Baoxing Huai, Wenhan Luo, Weihua Luo, Zheng Zhang, Baotian Hu, and Min Zhang. Perception, reason, think, and plan: A survey on large multimodal reasoning models, 2025.

[32] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[33] OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025.

[34] xAI. Grok 3 beta — the age of reasoning agents. https://x.ai/grok, 2025.

[35] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

[36] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[37] Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret. *arXiv preprint arXiv:2503.01491*, 2025.

[38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[39] John Schulman. Approximating kl divergence. http://joschu.net/blog/kl-approx.html, 2020.

[40] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[41] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[42] Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*, 2025.

[43] Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025. Accessed: 2025-01-24.

[44] Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling. *arXiv preprint arXiv:2502.06703*, 2025.

[45] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.

[46] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

[47] Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

[48] Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

[49] Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*, 2025.

[50] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

[51] Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.

[52] Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Guanjing Xiong, and Hongsheng Li. Ui-r1: Enhancing action prediction of gui agents by reinforcement learning. *arXiv preprint arXiv:2503.21620*, 2025.

[53] Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025.

[54] Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. *arXiv preprint arXiv:2504.14239*, 2025.

[55] Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. *arXiv preprint arXiv:2503.18470*, 2025.

[56] Kangrui Wang, Pingyue Zhang, Zihan Wang, Qineng Wang, Yaning Gao, Linjie Li, Zhengyuan Yang, Chi Wan, Hanyang Chen, Yiping Lu, and Manling Li. Vagen: Training vlm agents with multi-turn reinforcement learning, 2025. URL https://github.com/RAGEN-AI/VAGEN.

[57] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

[58] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[59] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[60] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[61] Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1358–1375, 2024.

[62] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

[63] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VOAMTA8jKu.

[64] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

[65] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[66] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

[67] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

[68] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

[69] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025.

[70] F Meng, L Du, Z Liu, Z Zhou, Q Lu, D Fu, B Shi, W Wang, J He, K Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[71] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[72] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[73] Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Hang Zhang, Yuming Jiang, Xin Li, Deli Zhao, Fan Wang, Yu Rong, Aixin Sun, and Shijian Lu. Mmr1: Advancing the frontiers of multimodal reasoning. https://github.com/LengSicong/MMR1, 2025.

[74] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[75] Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025.

[76] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

[77] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.

[78] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

[79] Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025.

[80] Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song, et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*, 2025.

[81] Wenyi Xiao, Leilei Gan, Weilong Dai, Wanggui He, Ziwei Huang, Haoyuan Li, Fangxun Shu, Zhelun Yu, Peng Zhang, Hao Jiang, et al. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*, 2025.

[82] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[83] Zhiqiang Wang, Pengbin Feng, Yanbin Lin, Shuzhang Cai, Zongao Bian, Jinghua Yan, and Xingquan Zhu. Crowdvlm-r1: Expanding r1 ability to vision language model for crowd counting using fuzzy group relative policy reward. *arXiv preprint arXiv:2504.03724*, 2025.

[84] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

[85] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*, 2025.

[86] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[87] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025.

[88] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.

[89] Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. Compile scene graphs with reinforcement learning. *arXiv preprint arXiv:2504.13617*, 2025.

[90] Lin Li, Wei Chen, Jiahui Li, and Long Chen. Relation-r1: Cognitive chain-of-thought guided reinforcement learning for unified relational comprehension. *arXiv preprint arXiv:2504.14642*, 2025.

[91] Biao Wang. R1-track: Direct application of mllms to visual object tracking via reinforcement learning. https://github.com/Wangbiao2/R1-Track, 2025.

[92] Kaibin Tian, Zijie Xin, and Jiazhen Liu. SeekWorld: Geolocation is a natural RL task for o3-like visual clue-tracking. https://github.com/TheEighthDay/SeekWorld, 2025. GitHub repository.

[93] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

[94] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.

[95] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[96] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.

[97] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.

[98] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025.

[99] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[100] Ya-Qi Yu, Minghui Liao, , Feilong Chen, Jihao Wu, and Chao Weng. R1-vision: Let's first take a look at the image. https://github.com/yuyq96/R1-Vision, 2025. Accessed: 2025-02-08.

[101] Lingfeng Ming, Yadong Li, Song Chen, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Ocean-r1: An open and generalizable large vision-language model enhanced by reinforcement learning. https://github.com/VLM-RL/Ocean-R1, 2025. Accessed: 2025-04-03.

[102] Xiaodong Wang and Peixi Peng. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video, 2025.

[103] Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. *arXiv preprint arXiv:2503.13377*, 2025.

[104] Hongyu Li, Songhao Han, Yue Liao, Jialin Gao, and Si Liu. Envolving temporal reasoning capability into lmms via temporal consistent reward. https://github.com/appletea233/Temporal-R1, 2025.

[105] Canhui Tang. Open llava-video-r1. https://github.com/Hui-design/Open-LLaVA-Video-R1, 2025. Accessed: 2025-03-18.

[106] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[107] Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv preprint arXiv:2503.24376*, 2025.

[108] Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*, 2025.

[109] Kun Ouyang. Spatial-r1: Enhancing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

[110] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[111] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025.

[112] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

[113] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

[114] Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv preprint arXiv:2504.20930*, 2025.

[115] Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, et al. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*, 2025.

[116] Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. *arXiv preprint arXiv:2504.12680*, 2025.

[117] Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025.

[118] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

[119] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

[120] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

[121] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.

[122] Zhiyuan Liu, Yuting Zhang, Feng Liu, Changwang Zhang, Ying Sun, and Jun Wang. Othink-mr1: Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning. *arXiv preprint arXiv:2503.16081*, 2025.

[123] Lujun Gui and Qingnan Ren. Training reasoning model with dynamic advantage estimation on reinforcement learning. https://github.com/ShadeCloak/ADORA, 2025.

[124] Shunyu Yao. The second half. https://ysymyth.github.io/The-Second-Half/, 2025.

[125] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

[126] Tianyu Pang Haonan Wang, Chao Du. V1: Toward multimodal reasoning by designing auxiliary tasks, 2025. URL https://v1-videoreasoning.notion.site.

[127] Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, et al. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*, 2025.

[128] Huanqia Cai, Yijun Yang, and Winston Hu. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *arXiv preprint arXiv:2502.00698*, 2025.

[129] Pengfei Zhou, Fanrui Zhang, Xiaopeng Peng, Zhaopan Xu, Jiaxin Ai, Yansheng Qiu, Chuanhao Li, Zhen Li, Ming Li, Yukang Feng, et al. Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models. *arXiv preprint arXiv:2504.05782*, 2025.

[130] Peijie Wang, Zhong-Zhi Li, Fei Yin, Xin Yang, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. *arXiv preprint arXiv:2502.20808*, 2025.

[131] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Pulsecheck457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. *arXiv e-prints*, pages arXiv–2502, 2025.

[132] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025.

[133] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

[134] Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *arXiv preprint arXiv:2504.14693*, 2025.

[135] Siqi Li, Yufan Shen, Xiangnan Chen, Jiayi Chen, Hengwei Ju, Haodong Duan, Song Mao, Hongbin Zhou, Bo Zhang, Pinlong Cai, et al. Gdi-bench: A benchmark for general document intelligence with vision and reasoning decoupling. *arXiv preprint arXiv:2505.00063*, 2025.

[136] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025.

[137] Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. ivispar–an interactive visual-spatial reasoning benchmark for vlms. *arXiv preprint arXiv:2502.03214*, 2025.

[138] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

[139] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[140] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 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.

[141] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[142] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

[143] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[144] Yan Ma, Steffi Chern, Xuyang Shen, Yiran Zhong, and Pengfei Liu. Rethinking rl scaling for vision language models: A transparent, from-scratch framework and comprehensive evaluation scheme. *arXiv preprint arXiv:2504.02587*, 2025.

[145] Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, et al. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*, 2025.

[146] Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, et al. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*, 2025.

[147] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025.

[148] Xiaokun Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, Yang Liu, et al. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning. *arXiv preprint arXiv:2505.07263*, 2025.

[149] Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. Mllm-protector: Ensuring mllm's safety without hurting performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16012–16027, 2024.

[150] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*, pages 388–404. Springer, 2024.

[151] Tianle Gu, Zeyang Zhou, Kexin Huang, Liang Dandan, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Yujiu Yang, Yan Teng, Yu Qiao, et al. Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:7256–7295, 2024.

[152] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Safety of multimodal large language models on images and text. In *IJCAI*, 2024.

[153] Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. The hidden risks of large reasoning models: A safety assessment of r1. *arXiv preprint arXiv:2502.12659*, 2025.

[154] Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*, 2025.

[155] Yang Yao, Xuan Tong, Ruofan Wang, Yixu Wang, Lujundong Li, Liang Liu, Yan Teng, and Yingchun Wang. A mousetrap: Fooling large reasoning models for jailbreak with chain of iterative chaos. *arXiv preprint arXiv:2502.15806*, 2025.

[156] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[157] Lilian Weng. Reward hacking in reinforcement learning. *lilianweng.github.io*, Nov 2024. URL https://lilianweng.github.io/posts/2024-11-28-reward-hacking/.

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

[159] Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 291–300, 2024.

[160] Xize Cheng, Zhengzhou Cai, Zehan Wang, Shengpeng Ji, Ziyue Jiang, Tao Jin, and Zhou Zhao. R1V-Free: Advancing Open-World Visual Reasoning with Label-Free AI Feedback, 2025. URL https://github.com/Exgc/R1V-Free.

A Summary of works done on RFT for MLLMs

A.1 Vision (Image)

Papers

[2505] [OpenThinkIMG [93]] OpenThinkIMG: Learning to Think with Images via Visual Tool Reinforcement Learning [🤗Model] [🤗Datasets] [💻Code]

[2505] [DanceGRPO (Gen) [120]] DanceGRPO: Unleashing GRPO on Visual Generation [🌐Project] [💻Code]

[2505] [Flow-GRPO (Gen) [119]] Flow-GRPO: Training Flow Matching Models via Online RL [🤗Models] [💻Code]

[2505] [X-Reasoner [145]] X-Reasoner: Towards Generalizable Reasoning Across Modalities and Domains [💻Code]

[2505] [T2I-R1 (Gen) [118]] T2I-R1: Reinforcing Image Generation with Collaborative Semantic-level and Token-level CoT [💻Code]

[2504] [FAST [81]] Fast-Slow Thinking for Large Vision-Language Model Reasoning [💻Code]

[2504] [Skywork R1V2 [80]] Skywork R1V2: Multimodal Hybrid Reinforcement Learning for Reasoning [🤗Models] [💻Code]

[2504] [Relation-R1 [90]] Relation-R1: Cognitive Chain-of-Thought Guided Reinforcement Learning for Unified Relational Comprehension [💻Code]

[2504] [R1-SGG [89]] Compile Scene Graphs with Reinforcement Learning [💻Code]

[2504] [NoisyRollout [79]] Reinforcing Visual Reasoning with Data Augmentation [🤗Models] [🤗Datasets] [💻Code]

[2504] [SimpleAR (Gen) [117]] SimpleAR: Pushing the Frontier of Autoregressive Visual Generation through Pretraining, SFT, and RL [🤗Models] [💻Code]

[2504] [VL-Rethinker [78]] VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning [🌐Project] [🤗Models] [🤗Dataset] [💻Code]

[2504] [Kimi-VL [99]] Kimi-VL Technical Report [🌐Project] [🤗Models] [🤗Demo] [💻Code]

[2504] [VLAA-Thinking [98]] SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models [🤗Models] [🤗Dataset] [💻Code]

[2504] [Perception-R1 [88]] Perception-R1: Pioneering Perception Policy with Reinforcement Learning [🤗Models] [🤗Datasets] [💻Code]

[2504] [SoTA with Less [97]] SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement [🤗Model] [🤗Datasets] [💻Code]

[2504] [VLM-R1 [82]] VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model [🤗Model] [🤗Dataset] [🤗Demo] [💻Code]

[2504] [CrowdVLM-R1 [83]] CrowdVLM-R1: Expanding R1 Ability to Vision Language Model for Crowd Counting using Fuzzy Group Relative Policy Reward [🤗Dataset] [💻Code]

[2504] [MAYE [144]] Rethinking RL Scaling for Vision Language Models: A Transparent, From-Scratch Framework and Comprehensive Evaluation Scheme [🤗Dataset] [💻Code]

[2503] [Q-Insight [87]] Q-Insight: Understanding Image Quality via Visual Reinforcement Learning [💻Code]

[2503] [Reason-RFT [96]] Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning [🌐Project] [🤗Dataset] [💻Code]

[2503] [OpenVLThinker [77]] OpenVLThinker: An Early Exploration to Vision-Language Reasoning via Iterative Self-Improvement [🤗Model] [💻Code]

[2503] [Think or Not Think [85]] Think or Not Think: A Study of Explicit Thinking in Rule-Based Visual Reinforcement Fine-Tuning [🤗Models] [🤗Datasets] [💻Code]

[2503] [OThink-MR1 [122]] OThink-MR1: Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning

[2503] [R1-VL [76]] R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization [🤗Model] [💻Code]

[2503] [Skywork R1V [80]] Skywork R1V: Pioneering Multimodal Reasoning with Chain-of-Thought [🤗Model] [💻Code]

[2503] [R1-Onevision [74]] R1-Onevision: Advancing Generalized Multimodal Reasoning through Cross-Modal Formalization [🤗Model] [🤗Dataset] [🤗Demo] [💻Code]

[2503] [VisualPRM [45]] VisualPRM: An Effective Process Reward Model for Multimodal Reasoning [🌐Project] [🤗Model] [🤗Dataset] [🤗Benchmark]

[2503] [LMM-R1 [72]] LMM-R1: Empowering 3B LMMs with Strong Reasoning Abilities Through Two-Stage Rule-Based RL [💻Code]

[2503] [Curr-ReFT [121]] Boosting the Generalization and Reasoning of Vision Language Models with Curriculum Reinforcement Learning [🤗Models] [🤗Dataset] [💻Code]

[2503] [VisualThinker-R1-Zero [84]] R1-Zero's "Aha Moment" in Visual Reasoning on a 2B Non-SFT Model [💻Code]

[2503] [Vision-R1 [71]] Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models [💻Code]

[2503] [Seg-Zero [86]] Seg-Zero: Reasoning-Chain Guided Segmentation via Cognitive Reinforcement [🤗Model] [🤗Dataset] [💻Code]

[2503] [MM-Eureka [70]] MM-Eureka: Exploring Visual Aha Moment with Rule-based Large-scale Reinforcement Learning [🤗Models] [🤗Dataset] [💻Code]

[2503] [Visual-RFT [95]] Visual-RFT: Visual Reinforcement Fine-Tuning [🌐Project] [🤗Data sets][💻Code]

[2501] [PARM++ (Gen) [46]] Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step [🌐Project] [🤗Model] [💻Code]

[2501] [Kimi k1.5 [32]] Kimi k1.5: Scaling Reinforcement Learning with LLMs [🌐Project]

[2501] [Virgo [69]] Virgo: A Preliminary Exploration on Reproducing o1-like MLLM [🤗Model] [💻Code]

[2412] [Mulberry [68]] Mulberry: Empowering MLLM with o1-like Reasoning and Reflection via Collective Monte Carlo Tree Search [🤗Model] [💻Code]

[2411] [Insight-V [94]] Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models [🤗Model] [💻Code]

[2411] [InternVL2-MPO [67]] Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization [🌐Project] [🤗Model] [💻Code]

Open-Source Projects (Repository without Paper)

[R1-V [140]] [💻Code] [🤗Datasets] [🎯Blog]

[Multimodal Open R1 [138]] [💻Code] [🤗Model] [🤗Dataset]

[MMR1 [73]] [💻Code] [🤗Model] [🤗Dataset]

[R1-Multimodal-Journey [70]] [💻Code]

[R1-Vision [100]] [💻Code] [🤗Cold-Start Datasets]

[Ocean-R1 [101]] [💻Code] [🤗Models] [🤗Datasets]

[R1V-Free [160]] [💻Code] [🤗Models] [🤗Dataset]

[SeekWorld [92]] [💻Code] [🤗Model] [🤗Dataset] [🤗Demo]

[R1-Track [91]] [💻Code] [🤗Models] [🤗Datasets]

A.2 Vision (Video)

Papers

[2504] [TinyLLaVA-Video-R1 [111]] TinyLLaVA-Video-R1: Towards Smaller LMMs for Video Reasoning [🤗Model] [💻Code]

[2504] [VideoChat-R1 [110]] VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning [🤗Model] [💻Code]

[2504] [Spatial-R1 [109]] Spatial-R1: Enhancing MLLMs in Video Spatial Reasoning [🤗Model] [🤗Datasets] [💻Code]

[2504] [R1-Zero-VSI [108]] Improved Visual-Spatial Reasoning via R1-Zero-Like Training [💻Code]

[2503] [SEED-Bench-R1 [107]] Exploring the Effect of Reinforcement Learning on Video Understanding: Insights from SEED-Bench-R1 [🤗Dataset] [💻Code]

[2503] [Video-R1 [106]] Video-R1: Reinforcing Video Reasoning in MLLMs [🤗Model] [🤗Data set][💻Code]

[2503] [TimeZero [103]] TimeZero: Temporal Video Grounding with Reasoning-Guided LVLM [🤗 Model] [💻Code]

Open-Source Projects (Repository without Paper)

[Open R1 Video [102]] [💻Code] [🤗Model] [🤗Dataset]

[Temporal-R1 [104]] [💻Code] [🤗Models]

[Open-LLaVA-Video-R1 [105]] [💻Code]

A.3 Medical Vision

Papers

[2504] [ChestX-Reasoner [114]] ChestX-Reasoner: Advancing Radiology Foundation Models with Reasoning through Step-by-Step Verification

[2503] [Med-R1 [113]] Med-R1: Reinforcement Learning for Generalizable Medical Reasoning in Vision-Language Models [🤗Model] [💻Code]

[2502] [MedVLM-R1 [112]] MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning [🤗Model]

A.4 Embodied Vision

Papers

[2504] [Embodied-R [116]] Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning in Foundation Models via Reinforcement Learning [💻Code]

[2503] [Embodied-Reasoner [95]] Embodied-Reasoner: Synergizing Visual Search, Reasoning, and Action for Embodied Interactive Tasks [🌐Project] [🤗Dataset][💻Code]

A.5 Multimodal Reward Model

Papers

[2505] [Skywork-VL Reward [148]] Skywork-VL Reward: An Effective Reward Model for Multimodal Understanding and Reasoning [🤗Model] [💻Code]

[2505] [UnifiedReward-Think [147]] Unified Multimodal Chain-of-Thought Reward Model through Reinforcement Fine-Tuning [🌐Project] [🤗Models] [🤗Datasets] [💻Code]

[2505] [R1-Reward [146]] R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning [🤗Model] [🤗Dataset] [💻Code]

A.6 Audio

Papers

[2504] [SARI [49]] SARI: Structured Audio Reasoning via Curriculum-Guided Reinforcement Learning

[2503] [R1-AQA [48]] Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering [🤗Model] [💻Code]

[2503] [Audio-Reasoner [47]] Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models [🌐Project] [🤗Model] [💻Code]

A.7 Omni

Papers

[2505] [EchoInk-R1 [51]] EchoInk-R1: Exploring Audio-Visual Reasoning in Multimodal LLMs via Reinforcement Learning [🤗Model] [🤗Dataset] [💻Code]

[2503] [R1-Omni [50]] R1-Omni: Explainable Omni-Multimodal Emotion Recognition with Reinforcement Learning [🤗Model] [💻Code]

A.8 GUI

Papers

[2504] [InfiGUI-R1 [54]] InfiGUI-R1: Advancing Multimodal GUI Agents from Reactive Actors to Deliberative Reasoners [🤗Model] [💻Code]

[2504] [GUI-R1 [52]] GUI-R1: A Generalist R1-Style Vision-Language Action Model For GUI Agents [🤗Model] [🤗Dataset] [💻Code]

[2503] [UI-R1 [53]] UI-R1: Enhancing Action Prediction of GUI Agents by Reinforcement Learning

A.9 Framework

Open-Source Project (Repository without Paper)

[EasyR1 [142]] [💻Code]

A.10 Metaverse

Paper

[2503] [MetaSpatial [55]] MetaSpatial: Reinforcing 3D Spatial Reasoning in VLMs for the Metaverse [🤗Dataset] [💻Code]

A.11 Agents

Open-Source Project (Repository without Paper)

[VAGEN [56]] [💻Code]