AdaTooler-V: Adaptive Tool-Use for Images and Videos

Chaoyang Wang[1,6] Kaituo Feng[1*] Dongyang Chen[2] Zhongyu Wang[3]
Zhixun Li[4] Sicheng Gao[7] Meng Meng[6†] Xu Zhou[6] Manyuan Zhang[1]
Yuzhang Shang[5] Xiangyu Yue[1†]

[1]MMLab, CUHK [2]THU [3]SJTU [4]DB Group, CUHK [5]UCF [6]Sangfor [7]JMU

Home: https://github.com/CYWang735/AdaTooler-V
HF: https://huggingface.co/AdaTooler-V

Abstract

Recent advances have shown that multimodal large language models (MLLMs) benefit from multimodal interleaved chain-of-thought (CoT) with vision tool interactions. However, existing open-source models often exhibit blind tool-use reasoning patterns, invoking vision tools even when they are unnecessary, which significantly increases inference overhead and degrades model performance. To this end, we propose AdaTooler-V, an MLLM that performs adaptive tool-use by determining whether a visual problem truly requires tools. First, we introduce AT-GRPO, a reinforcement learning algorithm that adaptively adjusts reward scales based on the Tool Benefit Score of each sample, encouraging the model to invoke tools only when they provide genuine improvements. Moreover, we construct two datasets to support training: AdaTooler-V-CoT-100k for SFT cold start and AdaTooler-V-300k for RL with verifiable rewards across single-image, multi-image, and video data. Experiments across twelve benchmarks demonstrate the strong reasoning capability of AdaTooler-V, outperforming existing methods in diverse visual reasoning tasks. Notably, AdaTooler-V-7B achieves an accuracy of 89.8% on the high-resolution benchmark V*, surpassing the commercial proprietary model GPT-4o and Gemini 1.5 Pro.

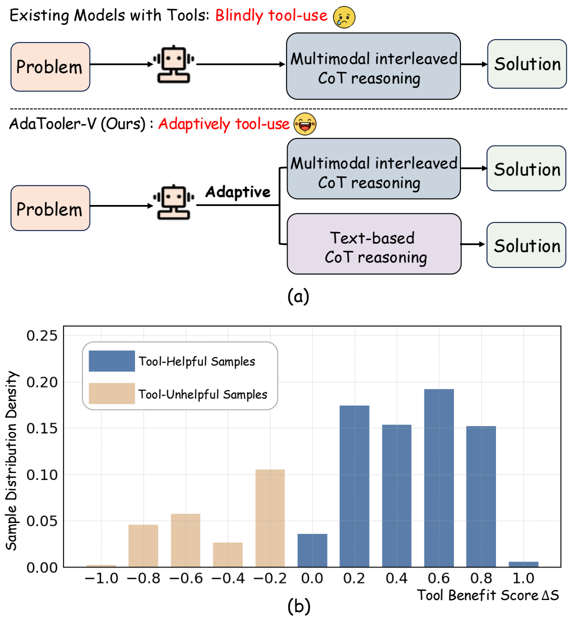


Figure 1: (a) Compared with existing models that blindly invoke vision tools, AdaTooler-V adaptively invokes tools by determining whether the problem truly requires tools. (b) Distribution of ΔS values in the AdaTooler-V-300k dataset, where positive and negative values correspond to tool-helpful and tool-unhelpful samples. Here, ΔS is computed as the difference in average accuracy when Qwen2.5-VL-72B-Instruct (Bai et al., 2025b) solves the same sample with and without tool-use.

1 Introduction

Recent advancements have highlighted the potential of rule-based Reinforcement Learning (RL) in enhancing the reasoning abilities of Large Language Models (LLMs) (Guo et al., 2025a; Yu et al., 2025; Zhang et al., 2025c). In particular, DeepSeek-R1 (Guo et al., 2025b) demonstrates the effectiveness of employing the GRPO (Shao et al., 2024) algorithm to incentivize strong reasoning with long Chain-of-Thought (CoT) in LLMs. Inspired by DeepSeek-R1's success, many subsequent studies have extended this paradigm to Multimodal Large Language Models (MLLMs) (Huang et al., 2025a; Tan et al., 2025; Wang et al., 2025b; Feng et al., 2025a; Fan et al., 2025b). Notable examples include Vision-R1 (Huang et al., 2025a), Video-R1 (Feng et al., 2025a) and OneThinker (Feng et al., 2025b), which apply RL to improve visual reasoning abilities.

In the field of multimodal reasoning, a rising trend is the multimodal interleaved CoT paradigm, also known as "Thinking with Images." In this paradigm, models dynamically interact with exter-

*Project Leader.
†Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 18058–18072

nal vision tools (e.g., cropping, frame extraction) throughout the reasoning process (Zheng et al., 2025a; Lai et al., 2025a; Su et al., 2025a; Zhang et al., 2025d). Such visual interactions enable the model to repeatedly focus on fine-grained visual details that text-only reasoning would otherwise overlook, thereby yielding substantial performance gains on challenging visual tasks.

However, existing models usually exhibit blind tool-use, invoking vision tools even when they are unnecessary. This phenomenon stems from a limitation in current approaches: models lack an explicit mechanism for determining when tools should be invoked, and reward functions may even blindly encourage tool-use. Nevertheless, as illustrated in Fig. 1, not all problems require tool-use. Many visual reasoning tasks can be solved efficiently using text-based CoT, and forcing tool-use can even degrade the final prediction quality. This is primarily because blind tool-use can induce overthinking (Fan et al., 2025a; Chen et al., 2024) during reasoning, driving the model to explore unnecessary trajectories, and deviate from the optimal reasoning path (Su et al., 2025b). Moreover, frequent and unnecessary tool invocations gradually weaken the model's reliance on the original visual input, making it harder for the model to focus on critical visual cues (Tian et al., 2025). In addition, blind tool-use may induce a series of meaningless tool operations (Li et al., 2025d). For tasks that inherently do not require tool-use, each extra tool-use introduces unnecessary computational overhead, thereby increasing the overall inference cost.

To address these challenges, we propose AdaTooler-V, an MLLM equipped with adaptive tool-use ability. Unlike previous approaches, AdaTooler-V adaptively adopts text-based CoT reasoning for problems that do not require tools, while progressively invoking vision tools to refine reasoning for tasks that do. The core of our approach is a novel reinforcement learning algorithm named Adaptive Tool-use GRPO (AT-GRPO). Specifically, we define a Tool Benefit Score ΔS for each sample, which quantifies the genuine performance gain provided by tool-use. AT-GRPO adaptively adjusts reward based on this score: it rewards tool-use only when it yields tangible improvements and penalizes redundant invocations. This mechanism enables the model to autonomously learn a favorable and generalizable reasoning strategy that optimizes both model performance and inference costs.

Besides, to support multimodal joint training, we construct two large-scale datasets: AdaTooler-V-CoT-100k for SFT cold start, and AdaTooler-V-300k for RL training. These datasets cover multiple modalities, including single-image, multi-image, and videos. They also span diverse visual reasoning tasks such as mathematics, visual counting, logical reasoning, spatial understanding, etc. Our two-phase training framework first establishes rich reasoning patterns and behavioral priors during the SFT stage using multi-round tool-interaction trajectories from AdaTooler-V-CoT-100k, and then further optimizes the model's reasoning strategy in the RL stage using AdaTooler-V-300k combined with the AT-GRPO algorithm. This enables AdaTooler-V to perform adaptive tool-use and achieve significant performance improvements over the base model across overall multimodal reasoning benchmarks. In summary, our contributions are as follows:

- We propose **AdaTooler-V**, an MLLM equipped with adaptive tool-use ability. To support training, we construct two datasets: **AdaTooler-V-CoT-100k** for SFT and **AdaTooler-V-300k** for RL training, covering diverse multimodal reasoning tasks and multiple modalities.
- We introduce **AT-GRPO**, a reinforcement learning algorithm that adjusts reward scales using a sample-specific Tool Benefit Score, ensuring tools are invoked only when they provide genuine improvements.
- Comprehensive experiments across 12 benchmarks demonstrate the effectiveness of AdaTooler-V. Notably, AdaTooler-V-7B achieves 89.8% accuracy on V* bench, outperforming the proprietary GPT-4o model.

2 Related Work

Multimodal Reasoning. Multimodal large language model reasoning aims to enable human-level inference across diverse modalities. Recent work, notably DeepSeek-R1 (Guo et al., 2025b), shows that RL-based post-training can substantially enhance LLM reasoning capabilities. Building upon the R1 paradigm, several subsequent works (Wang et al., 2025f; Yuan et al., 2025; Wang et al., 2025a; Li et al., 2025c) have applied similar post-training paradigms to multimodal large language models (MLLMs) to boost their performance across a variety of tasks. These include: Mathematical and

Figure 2: **Case reasoning trajectory of AdaTooler-V.** For single-image and video questions, the model alternates between internal reasoning, vision tool invocations and final answers, enabling zoom-in on fine-grained regions and inspection of informative clips. In contrast, for the multi-image clock example, AdaTooler-V solves the problem purely via text-based CoT, illustrating its ability to adaptively decide when vision tools are truly necessary.

scientific image-based visual question answering (VQA) (Peng et al., 2025; Huang et al., 2025b); Image segmentation and grounding (Liu et al., 2025a; Bai et al., 2025c; Shen et al., 2025; Liu et al., 2025b; Wang et al., 2025e,c; Yang et al., 2024); Video-based VQA (Feng et al., 2025a; Wang et al., 2025f; Li et al., 2025b; Cheng et al., 2025a). Unlike prior approaches that predominantly rely on text-based CoT, we adopt multimodal interleaved CoT, allowing the model to ground intermediate reasoning steps in visual observations and thereby enhance its visual understanding capabilities.

Thinking with Images. The "thinking with images" paradigm improves multimodal reasoning by allowing models to dynamically perform visual operations beyond text-based CoT, enabling iterative exploration and hypothesis verification (Su et al., 2025a,c,d; OpenAI, 2025; Zheng et al., 2025b; Zhang et al., 2025a). For example, OpenThinkIMG (Su et al., 2025c) introduces an end-to-end visual-tool reinforcement learning framework. MVoT (Li et al., 2025a) conceptualizes visualization as an intermediate representation within the reasoning process. PixelReasoner (Su et al., 2025a) leverages curiosity-driven reinforcement learning to incentivize pixel-level reasoning capabilities. Whereas, VITAL (Zhang et al., 2025a) explores incorporating multimodal interleaved CoT into video reasoning, thereby enhancing the model's video comprehension capabilities. Despite the remarkable progress of these approaches in multimodal reasoning, existing models often exhibit blind tool-use invocation during the reasoning process.

3 Method

3.1 Overview

Overall Agentic Pipeline. Given a user query and an input image/video, the policy model adaptively decides whether to invoke tools. For problems that don't require tool-use, the model can directly produce a single thought T to derive the final answer.

In contrast, when facing problems that require tool-use, the model follows an iterative thought–action–observation loop, sequentially generating thoughts T_i and actions C_i. Each action invokes image-related tools to obtain an observation E_i, which is appended to the history and fed back into the policy. This process continues until a final answer is produced or a predefined interac-

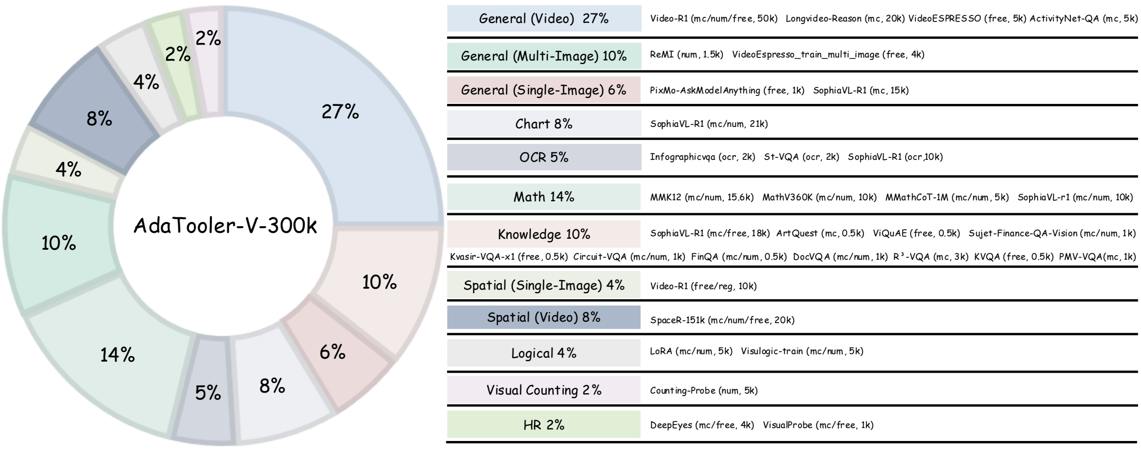


Figure 3: The data distribution of our AdaTooler-V-300k dataset.

tion limit is reached. The core components of the pipeline are detailed below.

- Thought (T_i): Represents the model's internal reasoning for action selection based on the interaction history and current observation, while encouraging diverse reasoning trajectories to support exploratory trial-and-error behavior.

- Action (C_i): The action space includes four primary vision tools: (1) *CropImg:* Zooms in or crops the image based on the specified bounding box. (2) *FrameAt:* Retrieves a single frame from the video at a specific time (in seconds). (3) *VideoClip:* Extracts a video clip between a start and end time; and (4) *PathTracer:* Draws a trajectory or connection between two points on the specified image. This formulation enables the model to act flexibly on any intermediate observation within the reasoning trajectory.

- Observation (E_i): The visual feedback resulting from executing C_i in the environment. Specifically, E_i corresponds to an image patch cropped from either the original input or a historical observation.

Two-Phase Training. Our training framework consists of two stages. (1) Supervised Fine-Tuning (SFT): the model is cold-started on multi-turn tool-interaction trajectories to learn coherent and diverse reasoning patterns. (2) Reinforcement Learning with Verifiable Rewards (RLVR): building on SFT, we apply AT-GRPO to move beyond rigid pattern matching and encourage the exploration of more effective reasoning strategies.

3.2 Training Data Collection

High-quality training data is essential for enhancing visual reasoning capabilities in MLLMs. In this section, we describe the construction of AdaTooler-V-300k for RL training and AdaTooler-V-CoT-100k for SFT cold-start.

Data Collection and Curation. The dataset covers single-image, multi-image, and video modalities. Image-based samples are designed to teach diverse reasoning skills across domains such as mathematics, spatial logic, and expert knowledge, enabling generalized reasoning in static settings. Video-based data focus on temporal reasoning, helping the model capture event dynamics and causal relationships over time. The dataset is constructed from multiple public sources with balanced sampling, and the final composition of AdaTooler-V-300k is summarized in Fig. 3, with details in Appendix A.

CoT Annotation. To facilitate effective initialization during the SFT stage, we leverage Qwen2.5-VL-72B-Instruct (Bai et al., 2025b) to automatically produce Chain-of-Thought (CoT) rationales for all samples in AdaTooler-V-300k dataset. The complete prompt specification employed for CoT generation is included in Appendix C. Following generation, we apply a sequence of rule-based filtering procedures to eliminate low-quality or semantically inconsistent outputs. This process yields a high-fidelity corpus, AdaTooler-V-CoT-100k, which forms the foundation for the cold-start stage of SFT.

Data Type and Rule-based Reward Design. Our RL framework adopts the rule-based reward

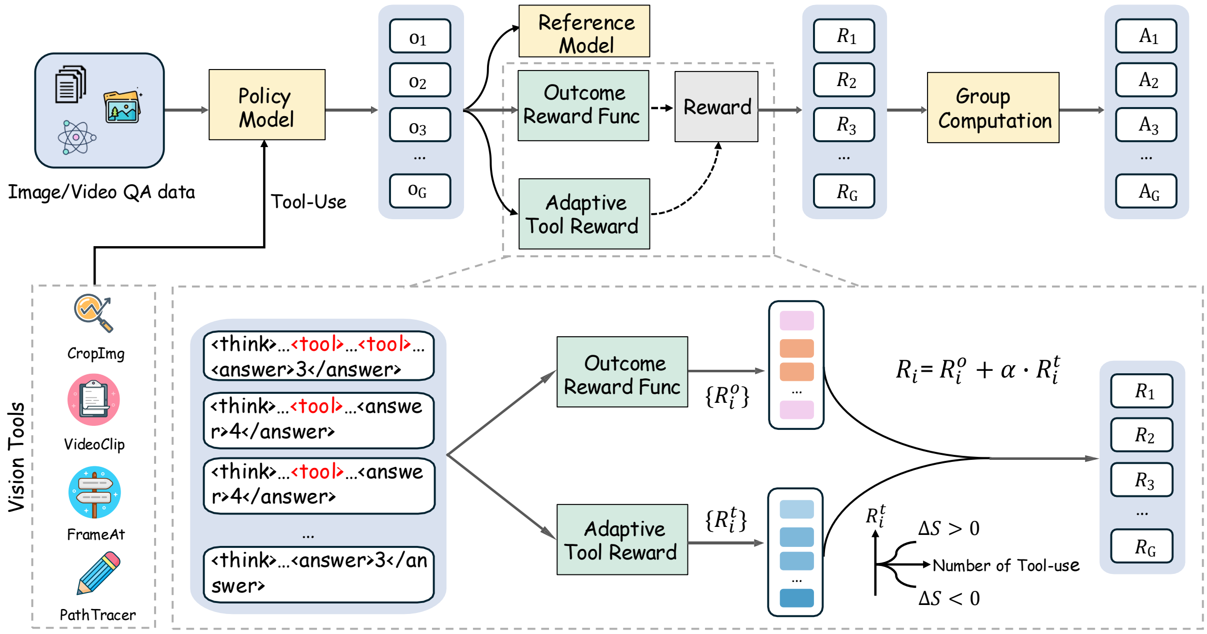


Figure 4: An illustration of our proposed AT-GRPO.

paradigm of DeepSeek-R1 (Guo et al., 2025b), requiring reliable and precise reward signals. Accordingly, most training samples use easily verifiable formats, such as multiple-choice and numerical QA, enabling stable RL training via simple rule-based rewards. To improve generalization, we also include a smaller portion of more complex data types, including free-form generation, OCR, and regression tasks.

The data types and their corresponding reward functions are summarized as follows: (1) **Multiple Choice:** Rewards are assigned based on an exact match between the model prediction and the ground-truth option. (2) **Numerical QA:** Rewards are given according to whether the predicted numerical value precisely matches the reference answer. (3) **OCR:** Rewards are computed using the Word Error Rate (WER), which measures the edit distance between the predicted text and the ground-truth transcription. (4) **Free-form QA:** Rewards are determined by the average of ROUGE-1, ROUGE-2, and ROUGE-L scores, assessing the similarity between the generated output and the reference answer.

3.3 Adaptive Tool-use GRPO Training

To enable adaptive tool-use during the reasoning process, we propose Adaptive Tool-use GRPO (AT-GRPO), which guides the model to invoke tools only when they offer a genuine performance gain, thereby improving model performance and reducing inference overhead, as illustrated in Fig. 4.

For each input query q_i, we define a Tool Benefit Score ΔS_i during the data annotation stage to quantify the performance improvement brought by tool-use:

$$\Delta S_i = S^+(q_i) - S^-(q_i) \qquad (1)$$

Here, S^+ and S^- denote the average accuracy of the model when reasoning with and without tool-use, respectively. Each query is evaluated eight times using Qwen2.5-VL-72B-Instruct (Bai et al., 2025b) (i.e., eight runs with tool-use and eight runs without tool-use) and the averaged accuracy gap is used as ΔS_i.

The adaptive-tool reward R_i^t is then formulated as:

$$R_i^t = \Delta S_i \cdot \exp\left(-\gamma \left(\frac{n_{\text{tool}} - n_{\text{max}}}{n_{\text{max}}}\right)^2\right) \qquad (2)$$

where ΔS_i represents the improvement in accuracy attributed to tool-use, n_{tool} is the number of tool-use during the reasoning trajectory, n_{max} denotes the maximum allowable number of tool-use, and γ controls the sensitivity of the Gaussian decay to tool frequency, making the reward variation smoother. Here we set $\gamma = 2$.

This design enables adaptive tool invocation by assessing whether a task truly requires visual tools. When tool use is unnecessary ($\Delta S_i < 0$), invoking tools is penalized with increasing cost, whereas

Table 1: Comparison of models on single-image and multi-image benchmarks. The first six evaluation benchmarks belong to single-image comprehension tasks, and the last two evaluation benchmarks belong to multi-image understanding tasks.

Model	V*	MME	InfoVQA	MMBench	MathVista	MMSI-Bench	SPAR-Bench
Proprietary Models							
GPT-4o (OpenAI, 2024)	65.2	2328	80.7	82.1	63.8	30.3	33.6
Gemini 1.5 Pro (Gemini Team, 2024)	71.7	–	81.0	–	63.9	36.9	–
Open-Source Models							
InternVL3-8B (Zhu et al., 2025)	–	2415.4	76.8	83.4	71.6	25.7	–
LLaVA-1.5-7B (Liu et al., 2024a)	–	1510.7	–	64.3	–	–	23.65
LLaVA-OneVision-7B (Li et al., 2024a)	–	1580.0	68.8	80.8	63.2	–	–
SophiaVL-R1-7B (Fan et al., 2025b)	–	2403.8	–	85.4	71.3	–	–
Qwen2.5-VL-7B-Instruct (Bai et al., 2025a)	78.5	2347.0	82.6	83.4	68.2	25.9	33.07
Open-Source o3-like Image Models							
Pixel Reasoner (Su et al., 2025a)	84.3	–	84.0	–	–	–	–
DeepEyes (Zheng et al., 2025a)	85.6	–	–	–	70.1	–	–
Mini-o3 (Lai et al., 2025b)	88.2	–	–	–	–	–	–
Thymes (Zhang et al., 2025d)	82.2	–	–	–	70.0	–	–
VILASR (Wu et al., 2025)	–	–	–	–	–	30.2	37.6
AdaTooler-V-7B	**89.8**	**2460.8**	**86.0**	**87.8**	**74.5**	**36.8**	**40.3**

when tool use is beneficial ($\Delta S_i > 0$), the model receives progressively higher rewards. We adopt this exponential form for its simplicity, differentiability, and stable gradient behavior during training.

The total reward for response i is defined as:

$$R_i = R_i^o + \alpha \cdot R_i^t \tag{3}$$

where α balances the relative weight of tool-use reward in the total objective and R_i^o denotes the base reward of response i, including correctness and formatting components, following the formulation in (Guo et al., 2025b). The combined reward R_i is then used to compute the advantage for policy optimization during GRPO training.

The advantage A_i is calculated within each group as:

$$A_i = \frac{R_i - \text{mean}(\{R_1, R_2, \ldots, R_G\})}{\text{std}(\{R_1, R_2, \ldots, R_G\})} \tag{4}$$

Following DeepSeek-R1 (Guo et al., 2025b), the final policy objective for AT-GRPO is given by:

$$\mathcal{J}_{\text{AT}-\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q),\, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(o|q)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right] \tag{5}$$

4 Experiments

4.1 Setup

Benchmarks. Following prior works (Feng et al., 2025a; Xiao et al., 2025; Wang et al., 2025d), we employ greedy decoding to systematically evaluate our proposed model and other baselines across a series of multimodal benchmarks. Specifically, for the image modality, we select seven benchmarks: V* (Wang et al., 2023), MME (Fu et al., 2025a), InfoVQA (Mathew et al., 2022), MMBench (Liu et al., 2024b), MathVista (Lu et al., 2024), MMSI-Bench (Yang et al., 2025b), and SPAR-Bench (Zhang et al., 2025b). For the video modality, we adopt five benchmarks: VSI-Bench (Yang et al., 2025a), VideoMMMU (Hu et al., 2025), MVBench (Li et al., 2024b), Video-MME (Fu et al., 2025b), and Video-Holmes (Cheng et al., 2025b).

Implementation Details. We use 8 NVIDIA H100 (80GB) GPUs to train our model. The training framework is based on verl-tool (Jiang et al., 2025), which extends the functionalities of verl (Sheng et al., 2024) and vLLM (Kwon et al., 2023), providing additional support for multimodal tool-augmented multi-turn training and evaluation. Our model is initialized based on Qwen2.5-VL-7B-Instruct (Bai et al., 2025a). First, we perform supervised fine-tuning (SFT) on the AdaTooler-V-CoT-100k dataset to obtain the Qwen2.5-VL-7B-SFT

Table 2: Comparison of models on video benchmarks.

Model	Frames	VSI-Bench	VideoMMMU	MVBench	Video-MME(w/o sub)	Video-Holmes
Proprietary Models						
GPT-4o (OpenAI, 2024)	–	34.0	61.2	64.6	71.9	42.0
Gemini 1.5 Pro (Gemini Team, 2024)	–	45.4	53.9	60.5	75.0	41.2
Open-Source Models						
InternVL3-8B (Zhu et al., 2025)	–	42.1	–	75.4	66.3	–
VideoChat-R1 (Li et al., 2025e)	–	–	–	67.9	72.2	33.0
Video-CCAM (Li et al., 2024b)	–	–	–	62.8	50.1	–
Video-XL (Shu et al., 2025)	–	–	52.3	55.3	55.5	–
Qwen2.5-VL-7B-Instruct (Bai et al., 2025a)	32	29.8	47.4	58.2	56.1	27.8
Qwen2.5-VL-7B-Instruct (Bai et al., 2025a)	64	30.9	49.1	59.8	58.6	29.9
Qwen2.5-VL-7B-Instruct (Bai et al., 2025a)	128	34.8	51.3	62.3	60.4	33.5
Video-R1 (Feng et al., 2025a)	–	37.1	52.4	64.8	61.4	36.5
Open-Source o3-like Video Models						
FrameMind (Ge et al., 2025)	–	–	–	64.2	60.9	–
Open-o3 Video (Meng et al., 2025)	–	–	52.3	–	63.6	–
Video-Thinker (Wang et al., 2025g)	–	–	–	–	–	43.2
VILASR (Wu et al., 2025)	–	45.4	–	–	–	–
AdaTooler-V-7B	32	46.7	54.6	68.4	62.5	55.6
AdaTooler-V-7B	64	47.9	55.1	70.2	63.4	56.4
AdaTooler-V-7B	128	**49.5**	**56.8**	**71.5**	**66.7**	**58.3**

model, where the number of epochs is set to 1, the batch size is set to 16, and the learning rate is set to 1×10^{-5}. Subsequently, we conduct reinforcement learning (RL) training on the AdaTooler-V-300k dataset to generate the final AdaTooler-V model, where the batch size is set to 32, the KL divergence coefficient to 0.04, and the learning rate to 5×10^{-7}. The maximum response length is limited to 4096 tokens. The model is optimized with AdamW (Loshchilov and Hutter, 2017) throughout the training process. The hyperparameter α in Eqn. 3 is set to 0.6.

4.2 Main Results

Image Benchmarks. As shown in Tab. 1, AdaTooler-V-7B achieves state-of-the-art performance on multiple single-image benchmarks. On the high-resolution V* benchmark, it reaches 89.8% accuracy, outperforming recent tool-based models and improving upon Qwen2.5-VL-7B-Instruct by +11.3%. The model also shows consistent gains on MME, MathVista, InfoVQA, and MMBench, with a notable 74.5% accuracy on MathVista, over 6 points higher than the base model, demonstrating strong cross-domain generalization. On multi-image reasoning tasks, AdaTooler-V-7B further achieves leading results on MMSI-Bench (36.8) and SPAR-Bench (40.3), highlighting its effectiveness in selectively invoking tools for complex spatial and relational reasoning across images.

Video Benchmarks. As is illustrated in Tab. 2, AdaTooler-V displays substantial performance gains over strong video-reasoning baselines. For example, our model achieves 46.7% on VSI-Bench, 54.6% on VideoMMMU, and 68.4% on MVBench using only 32 frames, surpassing both Qwen2.5-VL-7B-Instruct and Video-R1 based models. The Video-Holmes benchmark further highlights AdaTooler-V's strengths in complex, long-range video reasoning. Our method obtains 55.6%, compared to 27.8% for Qwen2.5-VL-7B-Instruct and 36.5% for Video-R1, showing more than a 2× improvement over the base model in causal, sequential inference settings. Moreover, we observe consistent performance gains across nearly all benchmarks as the number of input frames increases. This suggests that richer contextual cues and temporal information can further enhance the model's reasoning capability.

4.3 Ablation Study

4.3.1 Effectiveness of AT-GRPO

To validate the effectiveness of the AT-GRPO algorithm, we compare it with vanilla GRPO (without

Table 3: Ablation study on training stages.

Train Stage	V*	MathVista	VSI-Bench	MVBench	Avg.
GRPO	85.1	71.8	40.7	65.9	65.9
SFT+GRPO	87.0	73.2	42.3	67.7	67.6
SFT+AT-GRPO	89.8	74.5	46.7	68.4	69.9

Table 4: Ablation study on the α in Eqn. 3.

α	V*	MathVista	VSI-Bench	MVBench	Avg.
0.2	88.1	73.6	44.2	67.9	68.5
0.4	88.9	74.1	43.9	68.2	68.7
0.6	89.8	74.5	46.7	68.4	69.9
0.8	89.2	73.9	45.1	68.1	69.1

Table 5: Ablation study on tool-use.

Model	V*	MathVista	VSI-Bench	MVBench	Avg.
Qwen2.5-VL-7B	78.5	68.2	31.8	63.8	60.6
RL-wo-tool	84.4	72.6	39.9	65.0	65.5
AdaTooler-V-7B	89.8	74.5	46.7	68.4	69.9

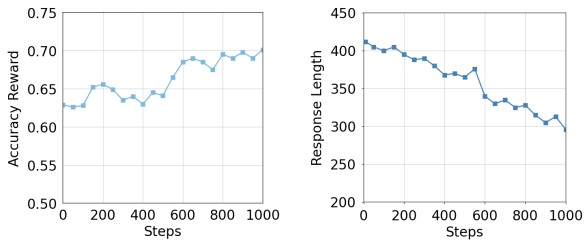
SFT cold start) and SFT+GRPO. As shown in Tab. 3, AT-GRPO yields clear performance gains, confirming that dynamically adjusting tool-use rewards via the Tool Benefit Score encourages necessary tool invocation while avoiding redundant interactions, resulting in more accurate reasoning.

4.3.2 Necessity of the SFT

We further investigate the necessity of the supervised fine-tuning (SFT) stage prior to reinforcement learning. As shown in the first row of Tab. 3, removing the SFT cold start results in clear performance degradation, as the model lacks structured priors for tool-based reasoning and struggles to generate coherent trajectories in early RL. In contrast, SFT provides essential tool-use and multimodal reasoning priors, leading to more stable training and allowing RL to more effectively refine adaptive reasoning behaviors.

4.3.3 Analysis of α

We further perform a analysis on the magnitude of the adaptive-tool reward, governed by the hyperparameter α, as illustrated in Tab. 4. We observe a moderate decrease in performance when α is set to 0.2 or 0.4, whereas α values of 0.6 and 0.8 yield overall comparable and consistently strong results. These findings suggest that the model exhibits a low sensitivity to the selection of α within a reasonable range.



(a) Accuracy Reward (b) Response Length

Figure 5: RL training curves.

4.3.4 Effectiveness of Tool-use

To assess the effectiveness of tool-use, we train a variant of Qwen2.5-VL-7B-Instruct using end-to-end RL with text-based CoT reasoning on the same training dataset (RL-wo-tool). As shown in Tab. 5, disabling tool interactions leads to consistent drops across four benchmarks. For example, from 89.8% to 84.4% on V* and from 46.7% to 39.9% on VSI-Bench. These results verify that tool-use provides complementary evidence beyond text-based reasoning and is essential for accurate multimodal understanding.

4.4 Training Curves

Fig. 5 presents the evolution of key metrics during RL training. As shown in Fig. 5(a), model accuracy steadily increases from approximately 0.60 to 0.70, indicating that reinforcement learning effectively enhances answer correctness. Fig. 5(b) shows that the average response length drops rapidly at early stages and then stabilizes. This is mainly because, although the model initially learns tool invocation during SFT, reinforcement learning encourages more efficient behaviors, leading the model to favor direct textual responses over unnecessary tool use for simpler samples.

5 Conclusion

We introduced AdaTooler-V, a multimodal large language model equipped with adaptive tool-use capability. To achieve this, we introduced AT-GRPO, a reinforcement learning algorithm that leverages a sample-specific Tool Benefit Score to dynamically modulate rewards, encouraging tools to be used only when they provide genuine performance gains. To support training, we curate two datasets, AdaTooler-V-CoT-100k for SFT cold start and AdaTooler-V-300k for RL. Experiments across twelve benchmarks validate the effectiveness of our approach. We believe it provides a foundation for future research on tool-augmented MLLMs.

Limitations

We discuss two main limitations of our work. First, our current study primarily focuses on visual modalities, including single-image and video-based reasoning tasks. While the proposed adaptive tool-use framework demonstrates strong effectiveness in these settings, it has not yet been validated on other modalities such as audio or audio-visual inputs. Extending the framework to support additional modalities would be a promising direction for future work, enabling more comprehensive multimodal reasoning capabilities. Second, the scope of tools considered in this work is mainly limited to vision-centric operations (e.g., image cropping and video frame extraction). Although these tools are sufficient for many visual reasoning tasks, they do not provide access to external world knowledge. In future work, our framework could be extended to incorporate more agentic multimodal tools, such as image search or web-based retrieval, allowing the model to actively acquire external knowledge and reason beyond the given perceptual inputs.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Grant No. 62306261), and The Shun Hing Institute of Advanced Engineering (SHIAE) Grant (No. 8115074). This study was supported in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. This work is also partially supported by Hong Kong RGC Strategic Topics Grant STG1/E-403/24-N, and CUHK-CUHK(SZ)-GDST Joint Collaboration Fund YSP26-4760949.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. 2025c. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025a. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025b. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. 2025a. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *arXiv preprint arXiv:2504.06514*.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. 2025b. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. *arXiv preprint arXiv:2505.17018*.

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

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, and 1 others. 2025b. Onethinker: All-in-one reasoning model for image and video. *arXiv preprint arXiv:2512.03043*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025a. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025b. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. 2025. Framemind: Frame-interleaved chain-of-thought for video reasoning via reinforcement learning. *arXiv preprint arXiv:2509.24008*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

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

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

Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, and 1 others. 2025. Verltool: Towards holistic agentic reinforcement learning with tool use. *arXiv preprint arXiv:2509.01055*.

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

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. 2025a. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv:2509.07969*.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. 2025b. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025a. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. 2025b. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*.

Hongyu Li, Manyuan Zhang, Dian Zheng, Ziyu Guo, Yimeng Jia, Kaituo Feng, Hao Yu, Yexin Liu, Yan Feng, Peng Pei, and 1 others. 2025c. Editthinker: Unlocking iterative reasoning for any image editor. *arXiv preprint arXiv:2512.05965*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Wenjun Li, Dexun Li, Kuicai Dong, Cong Zhang, Hao Zhang, Weiwen Liu, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025d. Adaptive tool use in large language models with meta-cognition trigger. *arXiv preprint arXiv:2502.12961*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

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

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and 1 others. 2025. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed 2025-09-19.

OpenAI. 2025. Thinking with images.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

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

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. 2025a. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. 2025b. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. *arXiv preprint arXiv:2505.00127*.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and 1 others. 2025c. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, and 1 others. 2025d. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*.

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

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. 2025. More thought, less accuracy? on the dual nature of reasoning in vision-language models. *arXiv preprint arXiv:2509.25848*.

Chaoyang Wang, Yangfan He, Yiyang Zhou, Yixuan Wang, Jiaqi Liu, Peng Xia, Zhengzhong Tu, Mohit Bansal, and Huaxiu Yao. 2025a. Knowing the answer isn't enough: Fixing reasoning path failures in lvlms. *arXiv preprint arXiv:2512.06258*.

Chaoyang Wang, Zeyu Zhang, Meng Meng, Xu Zhou, and Haiyun Jiang. 2025b. Vision-ekipl: External knowledge-infused policy learning for visual reasoning. *arXiv preprint arXiv:2506.06856*.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, and 1 others. 2025c. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*.

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

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and 1 others. 2025e. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. 2025f. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. 2025g. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*.

Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. 2023. Vstar: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. *arXiv preprint arXiv:2305.18756*.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2025. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*.

Zilin Xiao, Jaywon Koo, Siru Ouyang, Jefferson Hernandez, Yu Meng, and Vicente Ordonez. 2025. Proxythinker: Test-time guidance through small visual reasoners. *arXiv preprint arXiv:2505.24872*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, and 1 others. 2025b. Mmsibench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*.

Zhao Yang, Jiaqi Wang, Xubing Ye, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2024. Language-aware vision transformer for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and 1 others. 2025. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms. *arXiv preprint arXiv:2505.21327*.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. 2025a. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, and 1 others. 2025b. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. 2025c. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. *arXiv preprint arXiv:2506.03106*.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, and 1 others. 2025d. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025a. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025b. Deepeyes: Incentivizing “thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Dataset Distribution Details

The distribution of AdaTooler-V-300k dataset can be roughly categorized as follows:

- **General (Video, 81k):** A diverse collection of open-domain videos depicting everyday scenarios, designed to cultivate temporal comprehension and reasoning.
- **General (Multi-Image, 33k):** Multi-image reasoning tasks that test cross-view comparison and contextual integration.
- **General (Image, 18k):** General-purpose image question-answering data for foundational visual understanding.
- **Chart (Image, 24k):** Visual reasoning over charts, line graphs, and scientific figures, emphasizing data interpretation and quantitative logic.
- **OCR (Image, 15k):** Tasks requiring recognition and interpretation of embedded textual content, such as signs, forms, or documents.
- **Math (Image, 42k):** Image-based math reasoning problems, including formulas, geometric diagrams, and multi-step symbolic reasoning.
- **Knowledge (Image, 30k):** Visual commonsense and cross-disciplinary reasoning tasks to evaluate the integration of world knowledge with visual cues.
- **Spatial (Image, 12k):** Static spatial reasoning such as occlusion and positional inference.
- **Spatial (Video, 24k):** Focused on spatial reasoning in motion, including navigation, object tracking, and path planning, enhancing spatiotemporal understanding.
- **Logical (Image, 12k):** Visual logic tasks involving pattern recognition and rule-based reasoning.
- **Visual Counting (Image, 6k):** Object counting and density estimation for quantitative perception.
- **High-Resolution (Image, 6k):** Fine-grained visual understanding with small-object and texture recognition.

B Reasoning Examples

This section presents representative reasoning examples generated by AdaTooler-V-7B, as shown in Fig. 6 and Fig. 7.

C Prompt Template for Training and Inference

Fig. 8 illustrates the prompt template for training and inference of all models. We also use this prompt for the COT annotation.

Question: What kind of animal is in the poster?
(A) Dove (B) horse (C) dog (D) cat

AdaTooler-V-7B: <think>The image shows a modern church with a tall structure and stained glass windows. The poster on the church features an image of a bird flying, which is depicted as a dove.\n\nNow I will zoom in to get a clearer view of the poster.</think> <tool_call>{\"name\": \"crop_image\", \"arguments\": {\"bbox_2d\": [0.4, 0.5, 0.5, 0.6], \"target_image\": 1}}</tool_call>>

<think>The cropped image focuses on the upper part of the poster, confirming that the animal in the poster is indeed a dove.</think><answer>A</answer>

Figure 6: An example of AdaTooler-V-7B's reasoning output on V* Benchmark.

Question: What is the primary goal of the person in the video, as inferred from the interaction dynamics, textual cues, and video composition?

AdaTooler-V-7B: <think>The video's primary purpose is to promote a YouTube channel rather than simply showcase reptiles. On-screen questions and visual answers are used to maintain viewer interest, and the appearance of a clickable "Subscribe" button, notification prompts, and a full end screen with. </think><tool_call>{\"name\": \"get_video_clip_frame\", \"arguments\": {\"start_time\": 350.00, \"end_time\": 396.00}}</tool_call>

<think>Based on the selected frame, the only logical conclusion is to promote a YouTube channel focused on reptile care through direct calls to action.</think>
<answer> to promote a YouTube channel focused on reptile care through direct calls to action</answer>

Figure 7: An example of AdaTooler-V-7B's reasoning output on MVBench.

```
You are a helpful assistant.
# Tools
You may call one or more functions to assist with the user query.

You are provided with function signatures within <tools></tools> XML tags:
<tools>
{"type": "function", "function": {"name": "crop_image", "description": "Zoom in on the image based on the bounding box coordinates.", "parameters": {"type": "object", "properties": {"bbox_2d": {"type": "array", "description": "coordinates for bounding box of the area you want to zoom in. minimum value is 0 and maximum value is the width/height of the image.", "items": {"type": "number"}}, "target_image": {"type": "number", "description": "The index of the image to crop. Index from 1 to the number of images. Choose 1 to operate on original image."}}, "required": ["bbox_2d", "target_image"]}}}

{"type": "function", "function": {"name": "FrameAt", "description": "Get a single frame at a specific time from the video.", "parameters": {"type": "object", "properties": {"time": {"type": "number", "description": "Time (in seconds) of the frame to extract."}}, "required": ["time"]}}}

{"type": "function", "function": {"name": "VideoClip", "description": "Extract a video clip between start and end times.", "parameters": {"type": "object", "properties": {"t_start": {"type": "number", "description": "Start time (in seconds) of the clip."}, "t_end": {"type": "number", "description": "End time (in seconds) of the clip."}}, "required": ["t_start", "t_end"]}}}

{"type": "function", "function": {"name": "PathTracer", "description": "Plot movement or connections between two points on the specified image.", "parameters": {"type": "object", "properties": {"target_image": {"type": "number", "description": "The index of the image to crop. Index from 1 to the number of images. Choose 1 to operate on original image."}, "start_point_2d": {"type": "array", "description": "Starting point coordinates [x1, y1] of the path. minimum value is 0 and maximum value is the width/height of the image.", "items": {"type": "number"}}, "end_point_2d": {"type": "array", "description": "Ending point coordinates [x2, y2] of the path. minimum value is 0 and maximum value is the width/height of the image.", "items": {"type": "number"}}}, "required": ["start_point_2d", "end_point_2d", "target_image"]}}}
</tools>

For each function call, return a json object with function name and arguments within <tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>
```

Figure 8: Prompt template for training and inference.