OneThinker: All-in-one Reasoning Model for Image and Video

Kaituo Feng[1,2] **Manyuan Zhang**[2†] **Hongyu Li**[2] **Kaixuan Fan**[1,2] **Shuang Chen**[2]
Yilei Jiang[1,2] **Dian Zheng**[1,2] **Peiwen Sun**[1] **Yiyuan Zhang**[1] **Haoze Sun**[2]
Yan Feng[2] **Peng Pei**[2] **Xunliang Cai**[2] **Xiangyu Yue**[1‡]
[1]MMLab, CUHK [2]Meituan

Home: https://github.com/tulerfeng/OneThinker
HF: https://huggingface.co/OneThink

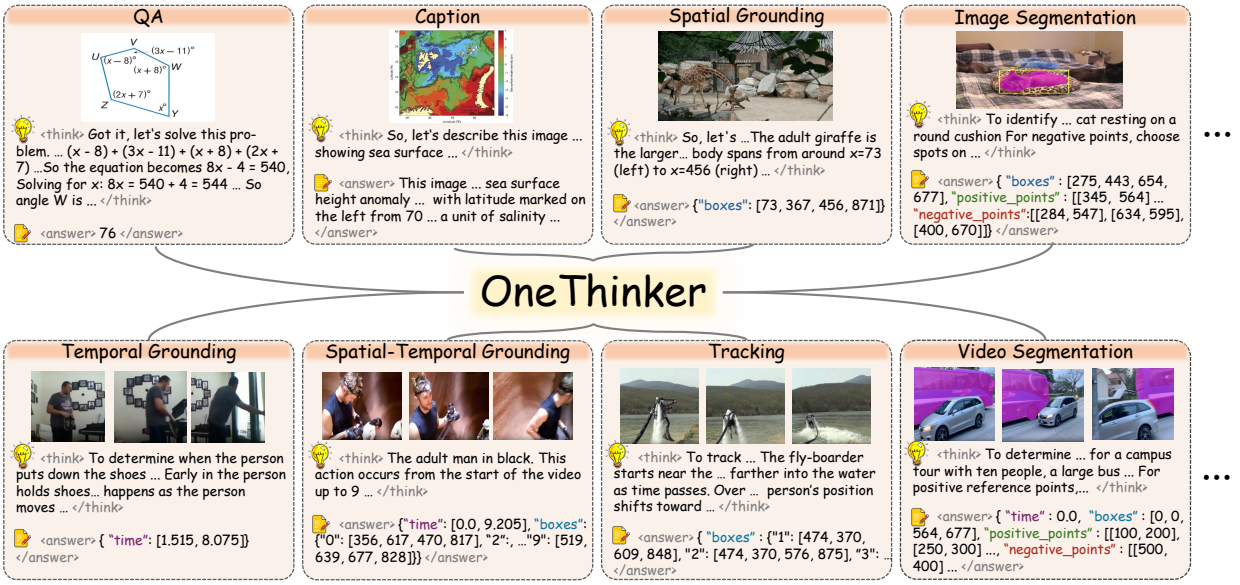


Figure 1: Overview of our OneThinker, which is capable of thinking across a wide range of tasks for image and video understanding.

ABSTRACT

Reinforcement learning (RL) has recently achieved remarkable success in eliciting visual reasoning within Multimodal Large Language Models (MLLMs). However, existing approaches typically train separate models for different tasks and treat image and video reasoning as disjoint domains. This results in limited scalability toward a multimodal reasoning generalist, which restricts practical versatility and hinders potential knowledge sharing across tasks and modalities. To this end, we propose OneThinker, an all-in-one reasoning model that unifies image and video understanding across diverse fundamental visual tasks, including question answering, captioning, spatial and temporal grounding, tracking, and segmentation. To achieve this, we construct the OneThinker-600k training corpus covering all these tasks and employ commercial models for CoT annotation, resulting in OneThinker-SFT-340k for SFT cold start. Furthermore, we propose EMA-GRPO to handle reward heterogeneity in multi-task RL by tracking task-wise moving averages of reward standard deviations for balanced optimization. Extensive experiments on diverse visual benchmarks show that OneThinker delivers strong performance on 31 benchmarks, across 10 fundamental visual understanding tasks. Moreover, it exhibits effective knowledge transfer between certain tasks and preliminary zero-shot generalization ability, marking a step toward a unified multimodal reasoning generalist. All code, model, and data are released.

[†]Project Leader.
[‡]Corresponding Author.

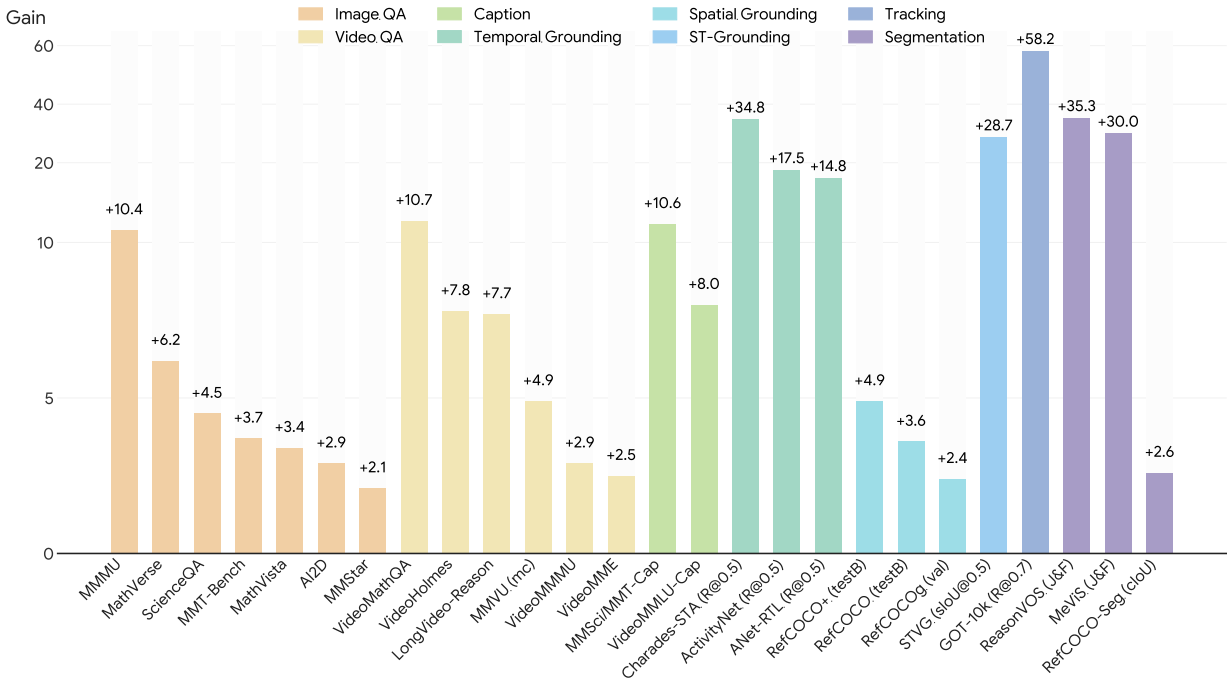


Figure 2: Performance gains of our model over Qwen3-VL-Instruct-8B across diverse visual tasks after training.

1 Introduction

Reasoning serves as a cornerstone in advancing Multimodal Large Language Models (MLLMs) toward artificial general intelligence (AGI), enabling them to perform step-by-step inference over complex visual–linguistic inputs [1, 2, 3]. Inspired by DeepSeek-R1 [4], a growing number of studies have witnessed the success of adopting reinforcement learning (RL) with the Group Relative Policy Optimization (GRPO) algorithm to enhance reasoning abilities [5, 6, 7, 8, 9]. For instance, Vision-R1 [10] and Video-R1 [9] demonstrate strong reasoning performance on image and video question answering, respectively, while VLM-R1 [11] excels in image detection and Seg-R1 [12] in segmentation. These advances underscore the remarkable effectiveness and broad potential of RL-based training for a wide range of visual tasks.

However, existing thinking models are usually designed to handle only a single task and operate exclusively on either images or videos. Such separation greatly limits their practical versatility and may also hinder the potential benefits of cross-task and cross-modal knowledge transfer. Although a few works have explored extending MLLMs with RL for multiple tasks [13, 14, 15], they are usually confined to limited subsets of visual tasks within a single modality. Furthermore, these approaches are often constrained by small-scale tuning, which limits their ability to generalize beyond specific domains. For instance, VideoChat-R1 [13] performs co-training on only three spatio-temporal perception tasks with merely 18k samples, and remains restricted to the video modality. Recognizing that vision inherently encompasses both static images and dynamic videos, and that real-world scenarios demand unified reasoning across diverse visual tasks, we pose a question:

Can we train an all-in-one multimodal reasoning generalist, which is capable of simultaneously handling both image and video understanding across diverse fundamental visual tasks?

To achieve this, we present OneThinker, a unified multimodal reasoning generalist capable of handling a wide range of visual reasoning tasks, including question answering, captioning, spatial and temporal grounding, tracking, and segmentation. First, we curate a large-scale dataset OneThinker-600k, comprising approximately 600k multimodal samples that jointly cover these fundamental visual tasks. We then employ a strong proprietary model Seed1.5-VL [16] to annotate and filter high-quality chain-of-thought (CoT) data, resulting in OneThinker-SFT-340k dataset for SFT cold start. Through joint multi-task training across both images and videos, OneThinker effectively learns to reason over spatial and temporal cues in a unified manner.

Besides, considering the distinct reward characteristics of heterogeneous visual tasks, we further introduce EMA-GRPO to improve RL training. This is motivated by two complementary imbalances: Standard GRPO suffers from intra-task imbalance because its sample-wise standard deviation (std) normalization favors low-variance rollouts [17, 18, 19, 20]; Conversely, removing this std normalization, as in Dr.GRPO [17], causes inter-task imbalance, where sparse-reward tasks (*e.g.*, math) dominate while dense ones (*e.g.*, detection) are suppressed. EMA-GRPO addresses both issues by maintaining task-wise exponential moving averages of reward standard deviations for normalization. This design allows each task to have a stable yet adaptive normalization scale that reflects its own reward dynamics. It balances intra-task weighting by reducing bias toward low-variance samples and prevents inter-task imbalance by using independent normalization statistics for different tasks, resulting in stable and balanced optimization across diverse visual tasks.

Extensive experiments demonstrate that OneThinker achieves consistently strong performance across diverse visual reasoning benchmarks. For example, OneThinker-8B reaches 70.6% accuracy on MMMU [21] and 64.3% on MathVerse [22] for image QA. In perception-oriented tasks such as grounding, tracking, and segmentation, our model also delivers strong results, for example, 84.4 R@0.5 on GOT-10k [23] and 54.9 J&F on ReasonVOS [24]. Moreover, unified training across tasks and modalities encourages effective knowledge sharing, allowing the model to transfer reasoning skills between several related tasks and exhibit preliminary zero-shot generalization on unseen scenarios.

Our main contributions are summarized as follows:

- We propose OneThinker, a unified multimodal reasoning generalist that handles a wide range of image and video tasks within a single model, including question answering, captioning, grounding, tracking, and segmentation. To support training, we construct the large-scale datasets OneThinker-600k and its CoT-annotated subset OneThinker-SFT-340k.
- To address the distinct reward characteristics of heterogeneous visual tasks, we introduce EMA-GRPO, which mitigates both intra-task and inter-task imbalance through task-wise adaptive normalization of reward statistics.
- Extensive experiments demonstrate that OneThinker achieves superior results on 31 benchmarks, across 10 fundamental visual understanding tasks. Besides, it promotes effective knowledge sharing in certain tasks, and exhibits preliminary zero-shot generalization abilities.

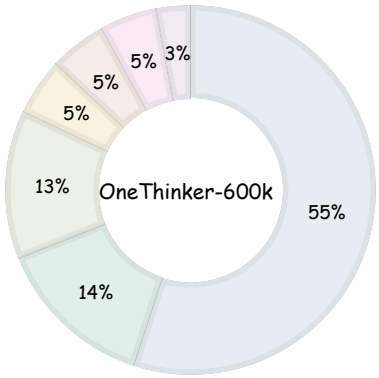
2 Related Works

2.1 Reinforcement Learning for LLM Reasoning

Reinforcement learning (RL) has emerged as a powerful technique for enhancing the reasoning capabilities of Large Language Models (LLMs) [7, 25, 26, 27, 28]. Recent studies, exemplified by DeepSeek-R1, adopt rule-based RL with Group Relative Policy Optimization (GRPO) [4] algorithm to directly optimize outcome-level rewards, enabling step-by-step reasoning without explicit intermediate supervision. The success of DeepSeek-R1 motivates a surge of works exploring this paradigm further [7, 29, 30]. For example, Dr.GRPO [17] introduces an unbiased optimization method that addresses the sample-wise standard deviation imbalance and response-length bias inherent in standard GRPO. Besides, GSPO [25] introduces a sequence-level RL algorithm that replaces token-wise ratios with sequence-level optimization, improving training stability for large-scale Mixture-of-Experts models. Critique-GRPO [7] integrates natural language feedback to guide policy optimization, enabling LLMs to refine their reasoning through critique-based self-improvement beyond standard RL fine-tuning. However, most existing research still focuses on single tasks or homogeneous reasoning objectives, where reward distributions remain relatively consistent.

2.2 Reasoning in MLLMs

Inspired by the success of reasoning in LLMs, a rising trend of works aims to bring this capability into MLLMs, enabling reasoning in different visual tasks [8, 31, 9, 32, 3, 33, 34, 35, 36]. For instance, Vision-R1 [10] tackles complex image reasoning in visual question answering, while Video-R1 [9] advances question answering over dynamic video inputs. Perception-R1 [14] and VLM-R1 [11] further extend this paradigm to image object detection, revealing the potential of RL for perception-oriented tasks. Seg-R1 [12] introduces a decoupled reasoning–segmentation framework that employs GRPO-based RL to generate explicit chain-of-thought reasoning and positional prompts for image segmentation tasks. Time-R1 [37] adapts RL-based post-training to temporal grounding in videos and achieves promising results, whereas VideoChat-R1 [13] applies reinforcement fine-tuning on three spatio-temporal tasks to enhance perception and reasoning



Rule-based QA 55%	LLaVA-Video (mc, 72k)	LongVila(mc, 29k)	SR-91k(mc/reg, 29k)	STAR(mc,11k)	UniGeo(mc, 3k)
ArxivQA(mc, 10k)	NEXT-QA(mc, 7k)	PTest(mc, 6k)	CLEVER(mc/num, 8k)	PMC-VQA(mc, 6k)	Spacellava(mc/reg, 7k)
TabMWP(mc/num, 4k)	Visulogic-train(mc, 4k)	AI2D-train(mc, 4k)	A-OKVQA(mc, 4k)	SciQA-train(mc, 4k)	GVLQA(mc, 4k)
FigureQA(mc, 2k)	OpenSpaces(mc/reg, 7k)	TQA(mc, 3k)	EXAMS-train(mc, 2k)	GeoQA+(mc, 2k)	Geometry3K(mc, 2k)
IconQA(mc, 2k)	Holmes-train(mc, 2k)	MapQA(mc, 1k)	Visual7W(mc, 1k)	VQA-RAD(mc, 0.4k)	GEOS(mc, 0.2k)
UniGeo(mc, 3k)	ARES(math, 55k)	MMK12(math, 16k)	IAM(ocr, 2k)	HME100k(ocr, 4k)	Chrome_writting(ocr, 2k)
TextVQA(ocr, 2k)	Rendered_text(ocr, 2k)	TextCaps(ocr, 2k)	TextOCR(ocr, 2k)	DVQA(num, 2k)	Super-CLEVR(num, 1k)
CLEVR-Math(num, 1k)	ChartQA(num, 1k)	PlotQA(num, 1k)			
Open-ended QA & Caption 13%	LLaVA-Video(7.5k)	OpenSpaces(4k)	Spacellava(3k)	Sharegpt4v(2k)	RoBUT SQA(2k)
VisualWebInstruct(6k)	ShareGPT4Video(29k)	VideoEspresso(10k)	Sharegpt4o(6k)	AI2D-train(4k)	Chart2text(2k)
Temporal Grounding 14%	ActivityNet-train(30k)	Vtg_it(15k)	DiDeMo(15k)	Queryd(15k)	Charades-train(12k)
HiREST(0.4k)					
Spatial Grounding 5%	RefCOCO(10k)	RefCOCOg(10k)	RefCOCO+(10k)		
Spatial-temporal Grounding3%	LLaVA-ST-STVG(19k)				
Tracking 5%	ElysiumTrack(10k)	TrackingNet(10k)	GOT-10k-train(10k)		
Segmentation 5%	Ref-SAV-train(8k)	ReVOS-train(5k)	RefCOCO 2014(4k)	RefCOCOg 2014(4k)	RefCOCO+ 2014(3k)
RefYT-VOS-train(3k)	MeViS-train(2k)	DAVIS-train(0.3k)			

Figure 3: Overview of our curated training dataset, including both image and video modalities for a diverse range of understanding tasks.

in video understanding. SophiaVL-R1 [38] introduces thinking-process rewards to improve RL training for image question answering. While these approaches have achieved remarkable progress in multimodal reasoning, most models remain restricted to limited tasks, and support either image or video reasoning alone.

3 Method

3.1 Dataset Construction

Data Collection and Curation. High-quality and diverse training data are essential for developing a unified multimodal reasoning generalist. To this end, we construct the **OneThinker-600k** corpus as the foundation for training, as illustrated in fig. 3. Our dataset covers both image and video modalities and spans a series of fundamental visual reasoning tasks, including rule-based QA, open-ended QA, captioning, spatial grounding, temporal grounding, spatio-temporal grounding, tracking, and segmentation. For perception-oriented tasks such as grounding, tracking, and segmentation, we require the model to output responses in a predefined JSON schema to ensure consistent formatting and enable automatic, verifiable reward computation. Details of prompts and formats are provided in the Appendix.

To ensure task diversity and balanced modality coverage, we collect data from a broad range of public training datasets and carefully curate samples across various domains and difficulty levels. The curated dataset is designed to equip the model with a broad spectrum of core reasoning abilities, such as logical reasoning, knowledge-based inference, spatial perception, temporal understanding, causal inference, *etc*. Together, these capabilities enable a unified multimodal reasoning generalist that can perform structured and coherent inference over both static and dynamic visual contexts.

CoT Annotation. To enable effective SFT initialization for reasoning, we leverage a strong proprietary model, **Seed1.5-VL** [16], to produce CoT annotations on the previously constructed OneThinker-600k corpus. For different tasks, we apply task-specific filtering thresholds to ensure the accuracy of retained CoT traces. After rule-based checking and quality validation, we obtain the CoT-annotated subset **OneThinker-SFT-340k**. This SFT dataset provides a diverse and reliable foundation for developing unified multimodal reasoning across a wide range of visual tasks.

3.2 Task Types and Rewards

All tasks are cast into a unified text interface, where the model first produces its internal reasoning inside `<think>...</think>` and then outputs a task-specific result inside `<answer>...</answer>`. For perception-oriented tasks, the `<answer>` block contains a structured representation (*e.g.*, time spans, bounding boxes, sparse points) following a predefined schema, which allows automatic parsing and verification. The overall reward is

$$R = R_{\text{acc}} \; + \; R_{\text{format}} \, , \tag{1}$$

where R_{acc} is task-specific accuracy reward and R_{format} is format reward. For tasks requiring structured outputs, R_{format} further checks whether the output follows the predefined schema.

Rule-based QA. This category includes multiple-choice, numerical, regression, math, and OCR tasks. For multiple-choice, numerical, and math problems, correctness is determined by whether the predicted and ground-truth answers are equivalent. Regression tasks are evaluated using the Mean Relative Accuracy (MRA) metric [39], which measures relative closeness between the prediction and the reference value across multiple tolerance levels. OCR tasks use the Word Error Rate to compute the reward. These rule-based tasks provide deterministic and interpretable feedback for discrete reasoning and quantitative prediction, forming a reliable foundation for reinforcement learning.

Open-ended QA & Caption. For open-ended question answering and captioning tasks, we employ an external reward model to provide a similarity score:

$$R_{\text{acc}} \;=\; \mathrm{RM}\big(q,\, \hat{a},\, a\big), \tag{2}$$

where q denotes the input query, $\hat{a}$ is the model-predicted answer, and a is the reference answer. In this work, we adopt POLAR-7B [40] as the reward model RM.

Temporal Grounding. Temporal grounding requires the model to identify the start and end time of the queried event in a video. The answer encodes a continuous time segment, and we measure accuracy using temporal IoU:

$$R_{\text{acc}} \;=\; \mathrm{tIoU}\big([\hat{s}, \hat{e}], [s, e]\big), \tag{3}$$

where $\mathrm{tIoU}(\cdot,\cdot)$ denotes the temporal intersection-over-union of two intervals. Here, $\hat{s}$ and $\hat{e}$ represent the predicted start and end timestamps, while s and e denote their corresponding ground-truth values.

Spatial Grounding. Spatial grounding requires the model to localize a target region by predicting a bounding box. The accuracy is measured using spatial intersection-over-union (sIoU) between predicted and ground-truth boxes:

$$R_{\text{acc}} \;=\; \mathrm{sIoU}\big(\hat{b}, b\big), \tag{4}$$

where $\hat{b}$ and b denote the predicted and ground-truth bounding boxes, respectively, and $\mathrm{sIoU}(\cdot,\cdot)$ represents their spatial overlap ratio.

Spatial-temporal Grounding. This task unifies temporal and spatial localization, requiring the model to predict both the temporal span of an event and the corresponding bounding boxes across frames. The accuracy is computed by combining temporal IoU and mean spatial IoU:

$$R_{\text{acc}} \;=\; \mathrm{tIoU}\big([\hat{s}, \hat{e}], [s, e]\big) \;+\; \overline{\mathrm{sIoU}}, \tag{5}$$

where $\hat{s}$ and $\hat{e}$ denote the predicted start and end times, and $\overline{\mathrm{sIoU}}$ represents the mean IoU between predicted and ground-truth boxes across frames.

Tracking. Tracking requires the model to predict a sequence of bounding boxes for a given target across video frames. The accuracy is measured as the mean IoU over all frames:

$$R_{\text{acc}} \;=\; \overline{\mathrm{sIoU}}, \tag{6}$$

where $\overline{\mathrm{sIoU}}$ is the averaged IoU between predicted and ground-truth bounding boxes throughout the trajectory.

Segmentation. Following prior works applying RL for image segmentation [12, 6, 41], the model predicts a bounding box along with a set of positive and negative points to identify target objects. These predictions are subsequently fed into SAM2 [42] to generate the final segmentation mask. For video segmentation, we further require the model to predict a keyframe time $\hat{t}$ indicating when the predicted boxes and points should be applied. Due to the high computational latency of running SAM2 on all rollouts for video segmentation, we omit the mask-based reward in this paper. All bounding boxes and point annotations are provided by Seed1.5-VL [16].

We define a Gaussian kernel $\mathcal{G}(d) = \exp\Big(-\frac{d^2}{2\sigma^2}\Big)$ that normalizes distances into $[0, 1]$. We set $\sigma = 50$ for spatial distances and $\sigma = 1$ for temporal distances.

For image segmentation, the accuracy reward combines bounding box overlap with Gaussian similarities over positive and negative point sets:

$$R_{\text{acc}} \;=\; \mathrm{sIoU}(\hat{b}, b) \;+\; \mathcal{G}(\mathrm{dis}_{+}) \;+\; \mathcal{G}(\mathrm{dis}_{-}), \tag{7}$$

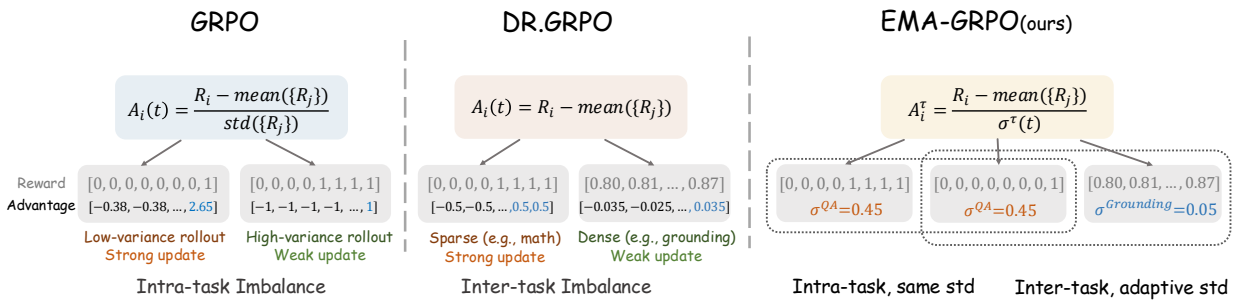


Figure 4: Comparison of advantage formulations in three RL algorithms.

where dis_+ denotes the minimum average distance between predicted and ground-truth positive points under optimal matching, and dis_- is defined similarly for negative points.

For video segmentation, a temporal Gaussian kernel is additionally applied to the predicted keyframe time:

$$R_{\text{acc}} = \text{sIoU}(\hat{b}, b) + \mathcal{G}(|\hat{t} - t|) + \mathcal{G}(\mathrm{dis}_+) + \mathcal{G}(\mathrm{dis}_-), \tag{8}$$

where $\hat{t}$ denotes the predicted keyframe timestamp and t is the annotated ground-truth time. In this paper, the number of positive points and negative points are both set to three.

3.3 EMA-GRPO

While GRPO has demonstrated strong capability in enhancing reasoning performance, its direct application to heterogeneous visual tasks would lead to biased optimization. We identify two complementary sources of imbalance that hinder effective multi-task training, as illustrated in fig. 4.

Intra-task Imbalance. Standard GRPO normalizes rewards within each prompt group by the group standard deviation to stabilize optimization. This normalization causes biased weighting among samples of the same task [17, 18, 19, 20]. Specifically, examples with very small or very large variance receive stronger updates, while medium-difficulty samples—whose rollouts usually have large variance—are under-optimized. As a result, the reinforcement learning within a task becomes biased.

Inter-task Imbalance. Conversely, removing the STD normalization as in Dr.GRPO [17] avoids the intra-task bias but introduces cross-task imbalance: different tasks vary in their reward scale and density, so sparse rewards (*e.g.*, math reasoning) dominate the optimization signal, whereas dense, small-range rewards (*e.g.*, grounding) are down-weighted. This imbalance causes the model to overfit a small subset of tasks and weakens its generalization across diverse visual reasoning settings.

EMA-based Normalization. To overcome both imbalances, we propose EMA-GRPO, which introduces task-wise adaptive normalization based on the exponential moving average (EMA) of reward statistics. For each task τ, we maintain EMA estimates of the first and second moments of its outcome rewards. Given the current batch of rewards $\{R_i\}$ belonging to task τ, let first-order moment $\mu^\tau(t) = \mathrm{mean}(\{R_i\})$ and second-order moment $\nu^\tau(t) = \mathrm{mean}(\{R_i^2\})$ at step t. We update the EMA moments as

$$\begin{aligned} m_1^\tau(t) &= \beta\, m_1^\tau(t-1) + (1-\beta)\, \mu^\tau(t), \\ m_2^\tau(t) &= \beta\, m_2^\tau(t-1) + (1-\beta)\, \nu^\tau(t), \end{aligned} \tag{9}$$

where β is the decay factor (set to 0.99). The task-wise standard deviation is then computed as

$$\sigma^\tau(t) = \sqrt{(m_2^\tau(t) - (m_1^\tau(t))^2}. \tag{10}$$

This moving statistic captures each task's intrinsic reward scale while adapting smoothly to changing reward distributions during training.

Then, the advantage in task τ is computed with its task-wise EMA standard deviation:

$$A_i^\tau(t) = \frac{R_i - \mathrm{mean}(\{R_j\})}{\sigma^\tau(t)}. \tag{11}$$

This adaptive normalization simultaneously resolves both intra-task and inter-task imbalance. Within each task, all rollouts share the same normalization scale $\sigma^\tau(t)$, which prevents the model from overemphasizing easy or hard samples while under-optimizing medium-difficulty ones. Across different tasks, each task maintains its own reward scale through an independent $\sigma^\tau(t)$, ensuring balanced gradient contributions regardless of differences in reward magnitude or density. For numerical stability during the initial stage, when $\sigma^\tau(t)$ has not yet stabilized, we clip the advantage to $[-5, 5]$. Together, these properties promote stable optimization and fair learning across heterogeneous visual reasoning tasks.

Training Objective. Following DeepSeek-R1, the final policy update adopts the standard GRPO objective with the EMA-normalized advantage:

$$\mathbb{E}_{q,\{o_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\Big(\min\Big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i^\tau(t),\ \text{clip}\Big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)},1-\epsilon,\,1+\epsilon\Big)A_i^\tau(t)\Big) - \beta_{\text{KL}}\,D_{\text{KL}}\big(\pi_\theta\,\|\,\pi_{\text{ref}}\big)\Big)\Bigg]. \tag{12}$$

The definition of variables and hyperparameters follows the standard GRPO [4].

4 Experiments

4.1 Setup

Training Details. Our model is trained on 32 NVIDIA H800 GPUs. In the SFT stage, we adopt Qwen-3-VL-Instruct-8B [43] as the base model and train it on our OneThinker-SFT-340k dataset. Subsequently, reinforcement learning is performed based on the SFT-initialized model using the OneThinker-600k corpus. For both SFT and RL, we sample image-video balanced sets for training. The batch size is set to 32 for SFT and 128 for RL. The learning rate is configured as 1×10^{-5} for SFT and 2×10^{-6} for RL, both optimized with AdamW. For efficiency, the maximum number of video frames is capped at 128 during training. The decay factor β is set to 0.99, following the common practice in EMA. The group size for EMA-GRPO is set to 8, and the KL regularization coefficient β_{KL} is fixed at 0.01. The maximum response length is limited to 4096 tokens. We discard rollouts that are entirely correct or incorrect during RL training, following the practice in [27]. Overall, the complete training process takes approximately 10 days.

Benchmarks. For evaluation, we adopt a variety of benchmarks corresponding to different visual reasoning tasks, covering question answering, captioning, spatial and temporal grounding, tracking, and segmentation, as presented in the experimental tables. For Qwen3-VL-Instruct, we report our reproduced results. We evaluate models using greedy decoding, following prior works [44, 9, 45].

4.2 Main Results

We evaluate OneThinker across a wide range of visual reasoning benchmarks covering both image and video modalities, as summarized in table 1, table 2, table 3, table 4, table 5, table 6, table 7, and table 8. Across all benchmarks, OneThinker demonstrates substantial improvements, showcasing its unified and transferable reasoning ability across tasks and modalities. Examples of reasoning responses for each task can be found in Appendix.

Image QA. OneThinker-8B consistently achieves top-tier performance for image QA across a diverse set of tasks spanning general knowledge, mathematics, science, and multimodal reasoning. Compared with strong open-source models such as Vision-R1-7B, VAPO-Thinker-7B, and Qwen3-VL-Instruct-8B, our model attains superior results on these benchmarks. For example, OneThinker reaches 70.6% on MMMU, 77.6% on MathVista, 64.3% on MathVerse, and 70.6% on MMStar, consistently outperforming all prior open-source competitors. These results demonstrate that our unified reasoning framework can effectively generalize to a wide range of complex image QA scenarios.

Video QA. In video QA, OneThinker-8B shows strong superiority over video-focused reasoning models. Across benchmarks including VideoMMMU, MMVU(mc), VideoMME, VideoHolmes, LongVideoBench, LongVideo-Reason, and VideoMathQA, OneThinker consistently ranks among the top performers. For instance, it achieves 66.2% on VideoMMMU, 70.5% on MMVU(mc), and 66.5% on VideoMME, outperforming specialized video reasoning models

Table 1: Performance of different models on image question answering benchmarks. For Qwen3-VL-Instruct-8B, we report our reproduced results under the same setting.

Models	Image QA							
	MMMU [21]	MathVista [46]	MathVerse [22]	MMBench [47]	MMStar [48]	ScienceQA [49]	AI2D [50]	MMT-Bench [51]
GPT-4o [52]	70.7	63.8	41.2	84.3	65.1	90.1	84.9	67.7
Gemini 2.5 Pro [53]	81.7	82.7	-	90.1	77.5	-	88.4	-
Seed1.5-VL [16]	77.9	85.6	-	89.9	77.8	-	87.3	
SophiaVL-R1-7B [38]	61.3	71.3	48.8	85.4	66.7	90.9		62.7
Vision-R1-7B [10]	-	73.5	52.4	-	-	-	-	-
MM-Eureka-7B [54]	57.3	73.0	50.3	-	64.4		-	-
VL-Rethinker-7B [44]	56.7	74.9	54.2	-	62.7	-	-	-
VAPO-Thinker-7B [55]	60.2	75.6	53.3	-	63.0	-	-	-
Qwen3-VL-Instruct-8B [43]	60.2	74.2	58.1	85.1	68.5	92.0	82.3	64.1
OneThinker-8B	**70.6**	**77.6**	**64.3**	**86.6**	**70.6**	**96.5**	**85.2**	**67.8**

Table 2: Performance of different models on video question answering benchmarks. For Qwen3-VL-Instruct-8B, we report our reproduced results under the same setting.

Models	Frames	Video QA						
		VideoMMMU[56]	MMVU(mc)[57]	VideoMME[58]	VideoHolmes[59]	LongVideoBench[60]	LongVideo-Reason[61]	VideoMathQA[62]
GPT-4o [52]		61.2	75.4	71.9	42.0	66.7	-	20.2
Gemini 2.5 Pro [53]	-	83.6	-	84.3	45.0	-	-	-
Seed1.5-VL [16]	-	81.4	-	77.9	-	74.0	-	-
VideoLLaMA3-7B [63]	-	-	-	66.2	-	59.8	-	-
InternVideo2.5-8B [64]	-	-	-	65.1	-	60.6	-	25.2
VideoChat-R1-7B [13]	-	46.4	-	60.0	33.0	-	67.2	27.6
LongVILA-R1-7B [61]	-	51.0	-	65.1	-	58.0	72.0	23.6
Video-R1-7B [9]	-	52.4	64.2	61.4	36.5	-	68.1	21.4
Qwen3-VL-Instruct-8B [43]	128	63.3	65.6	64.0	40.9	61.5	71.5	24.3
OneThinker-8B	128	**66.2**	**70.5**	**66.5**	**48.7**	**61.7**	**79.2**	**35.0**

such as VideoChat-R1-7B, VideoLLaMA3-7B, and InternVideo2.5-8B. Most notably, OneThinker obtains 79.2% on LongVideo-Reason, substantially surpassing Video-R1-7B (67.2%) and Qwen3-VL-Instruct-8B (71.5%). On VideoMathQA, a challenging video reasoning benchmark, OneThinker also leads all open-source models with a score of 35.0%. These results collectively verify that the effectiness of our proposed framework.

Image and Video Caption. On caption benchmarks, OneThinker maintains competitive or superior performance on both image and video captioning. For image captioning, it achieves 25.7 on MMSci-Caption and 57.9 on MMT-Caption, markedly outperforming Qwen3-VL-Instruct-8B (15.1 and 47.3 respectively) and significantly improving over LLaVA-1.5-7B. In video captioning, OneThinker reaches 28.0 on VideoMMLU-Caption, demonstrating effective video caption ability. This unified captioning ability reflects the model's strong visual descriptive skills.

Temporal Grounding. On temporal grounding tasks, OneThinker generally shows substantial improvements over existing temporal localization models. For example, on Charades, OneThinker-8B achieves performance that is

Table 3: Performance of different models on caption benchmarks.

Models	Frames	Image Caption		Video Caption
		MMSci-Caption[65]	MMT-Caption[51]	VideoMMLU-Caption[66]
GPT-4o [52]	-	27.0	-	53.9
LLaVA-1.5-7B [67]	-	11.8	-	22.3
Qwen3-VL-Instruct-8B [43]	128	15.1	47.3	20.0
OneThinker-8B	128	**25.7**	**57.9**	**28.0**

Table 4: Performance on temporal grounding benchmarks.

Models	Frame	Charades [68]				ActivityNet [69]				ANet-RTL [70]			
		R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
VTimeLLM [71]	-	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4	-	-	-	-
TimeSuite [72]	-	69.9	48.7	24.0	-	-	-	-	-	-	-	-	-
VideoChat-R1 [13]	-	83.1	**72.8**	**51.5**	**61.3**	50.4	32.2	16.2	34.3	-	-	-	-
Temporal-RLT [73]	-	80.2	68.3	44.5	57.9	56.9	38.4	20.2	39.1	40.2	22.7	10.9	26.3
Time-R1 [37]	-	78.1	60.8	35.3	-	58.6	39.0	21.4	-	-	-	-	-
Qwen3-VL-Instruct-8B [43]	128	58.0	33.5	13.1	36.7	39.9	26.1	15.3	29.1	36.2	27.5	18.3	26.6
OneThinker-8B	128	**83.5**	68.3	45.3	59.9	**65.0**	**43.6**	**25.7**	**45.9**	**62.0**	**42.3**	**22.7**	**43.2**

Table 5: Performance on spatial grounding benchmarks.

Models	RefCOCO[74]			RefCOCO+[74]			RefCOCOg[75]	
	testA	testB	val	testA	testB	val	test	val
Perception-R1 [14]	91.4	84.5	89.1	86.8	74.3	81.7	85.4	85.7
VLM-R1 [11]	-	-	90.5	-	-	84.3	-	87.1
DeepEyes [76]	-	-	89.8	-	-	83.6	-	86.7
Qwen3-VL-Instruct-8B [43]	92.2	85.3	89.9	89.6	77.8	84.5	86.7	86.8
OneThinker-8B	**93.7**	**88.9**	**92.0**	**91.4**	**82.7**	**87.0**	**88.8**	**89.2**

comparable to or better than previous models. On ActivityNet, our model attains 65.0 R@0.3, 43.6 R@0.5, and 25.7 R@0.7, confirming superior temporal grounding abilities. Furthermore, on the ANet-RTL benchmark, OneThinker achieves the best mIoU (43.2) among listed models. These results demonstrate the model's robust ability to reason about when events happen and accurately understand the fine-grained temporal information.

Spatial Grounding. For spatial grounding, OneThinker-8B also demonstrates state-of-the-art localization ability across the widely-used RefCOCO, RefCOCO+, and RefCOCOg benchmarks. In RefCOCO testA/testB/val sets, it achieves 93.7 / 88.9 / 92.0, outperforming prior strong models such as Perception-R1, DeepEyes, and Qwen3-VL-Instruct-8B. On RefCOCO+, OneThinker again leads with 91.4 / 82.7 / 87.0, consistently surpassing prior baselines by a large margin. On the more challenging RefCOCOg benchmark, the model achieves 88.8 / 89.2 (test/val), showing strong comprehension of long and descriptive referring expressions. These results highlight the model's strong spatial grounding abilities.

Spatial-Temporal Grounding. On spatial-temporal grounding tasks, which require simultaneous localization in both space and time, OneThinker delivers substantial improvements over previous systems. On the STVG benchmark, it achieves 34.9 tIoU@0.5, 39.5 tIoU, 40.3 sIoU@0.5, and 36.7 sIoU, outperforming Grounded-VideoLLM and Qwen3-VL-Instruct-8B by a large margin. Such gains emphasize OneThinker's capability to jointly reason about where and when events occur, even in complex videos involving multiple objects and temporal transitions.

Table 6: Performance on spatial-temporal grounding benchmarks.

Models	Frame	STVG [77]			
		tIoU@0.5	tIoU	sIoU@0.5	sIoU
GroundingGPT [78]	-	7,1	12.2	2.9	9.2
VTimeLLM [71]	-	7.1	15.5	-	-
Grounded-VideoLLM [79]	-	30.0	33.0	-	-
Qwen3-VL-Instruct-8B [43]	128	24.4	25.4	11.6	13.6
OneThinker-8B	128	**34.9**	**39.5**	**40.3**	**36.7**

Table 7: Performance on tracking benchmarks.

Models	GOT-10k [23]			
	AO	R@0.3	R@0.5	R@0.7
R1-Track [80]	68.0	-	76.6	-
VideoChat-R1 [13]	42.5	-	30.6	3.9
Qwen3-VL-Instruct-8B [43]	33.7	51.1	28.9	10.6
OneThinker-8B	**73.0**	**93.9**	**84.4**	**68.8**

Table 8: Performance on segmentation benchmarks. For RefCOCO series, we report results on the val set.

Models	Image Segmentation			Video Segmentation					
	RefCOCO [74]	RefCOCO+ [74]	RefCOCOg [75]	MeViS [81]			ReasonVOS [24]		
	cIoU	cIoU	cIoU	J	F	J&F	J	F	J&F
PixelLM-7B [82]	73.0	66.3	69.3	-	-	-	-	-	-
LISA-7B [83]	74.1	62.4	66.4	-	-	-	29.1	33.1	31.1
VISA-13B [84]	72.4	59.8	65.5	-	-	44.5	-	-	-
Seg-R1-7B [12]	74.3	62.6	**71.0**	-	-	-	-	-	-
ReferFormer [85]	-	-	-	29.8	32.2	31.0	30.2	35.6	32.9
VideoLISA-3.8B [24]	-	-	-	41.3	47.6	44.4	45.1	49.9	47.5
Qwen3-VL-Instruct-8B [43] + SAM2 [42]	73.2	66.2	68.3	19.4	26.4	22.9	16.6	22.7	19.6
OneThinker-8B	**75.8**	**67.1**	70.8	**48.8**	**56.7**	**52.7**	**51.1**	**58.7**	**54.9**

Tracking. As for tracking tasks, OneThinker reaches a high 73.0 AO, 93.9 R@0.3, 84.4 R@0.5, and 68.8 R@0.7 on GOT-10k, outperforming previous models like R1-Track and VideoChat-R1. Notably, our evaluation uses 32 frames for prediction, which is substantially more challenging than the 8-frame setting adopted by prior work VideoChat-R1. This large improvement illustrates that the unified reasoning architecture also yields strong single-object tracking capabilities, enabling reliable localization over long temporal sequences.

Image and Video Segmentation. OneThinker achieves the highest mean performance across both image and video segmentation benchmarks. For image segmentation (RefCOCO / RefCOCO+ / RefCOCOg), it obtains 75.8 / 67.1 / 70.8 cIoU on the val set, significantly outperforming PixelLM-7B, LISA-7B, VISA-13B, Seg-R1-7B. For video segmentation, it reaches 48.8 J, 56.7 F, and 52.7 J&F on MeViS, surpassing all previous open-source models. On ReasonVOS, it delivers 51.1 J, 58.7 F, and 54.9 J&F, again achieving the best performance across all competitors. These results demonstrate the model's fine-grained visual understanding ability in both static and dynamic environments.

Overall, OneThinker serves as a unified multimodal reasoning generalist that achieves strong performance across all major visual understanding tasks, showing strong potential for scalable and generalizable visual reasoning.

4.3 Ablation Study

In this section, we design three variants of OneThinker to verify the effectiveness of different components in our framework: (1) OneThinker-8B-SFT, which is trained only with SFT without RL; (2) OneThinker-8B-GRPO, which replaces our proposed EMA-GRPO with the original GRPO algorithm; (3) OneThinker-8B-DrGRPO, which adopts the Dr.GRPO [17] algorithm for RL training.

As shown in table 9, all ablated variants perform worse than OneThinker-8B across all tasks. Compared with the SFT baseline, RL consistently improves performance, demonstrating its effectiveness across diverse visual tasks. Replacing EMA-GRPO with standard GRPO or DrGRPO results in noticeable degradation, demonstrating the importance of addressing the intra-task imbalance and inter-task imbalance issues. Overall, This ablation study confirms the effectiveness of our unified RL framework and the proposed EMA-GRPO algorithm.

Table 9: Ablation study. Results are averaged over benchmarks within each task for comparison. For tasks with multiple evaluation metrics, we report the following: mIoU for temporal grounding, the mean of tIoU and sIoU for spatio-temporal grounding, AO for tracking, and the mean of cIoU (image) and J&F (video) for segmentation.

Models	QA	Temporal Grounding	Spatial Grounding	S.-T. Grounding	Tracking	Segmentation
Qwen3-VL-Instruct-8B	65.0	30.8	86.6	19.5	33.7	50.0
OneThinker-8B-SFT	67.0	31.8	87.8	27.1	48.1	62.8
OneThinker-8B-GRPO	67.2	46.9	86.5	34.5	65.5	62.3
OneThinker-8B-DrGRPO	67.6	46.3	88.2	34.0	67.8	61.2
OneThinker-8B	69.8	49.7	89.2	38.1	73.0	64.2

Table 10: Benefits between tasks and modalities analysis.

Variants	Image QA	Video QA	Tracking	Segmentation
OneThinker-wo-spatial-grounding	76.6	60.3	71.0	62.9
OneThinker-wo-temporal-grounding	77.2	59.5	67.2	63.3
OneThinker-wo-ImageQA	-	58.2	72.3	63.9
OneThinker	77.4	61.1	73.0	64.2

4.4 Benefits of Unified Training Analysis

To further investigate the potential benefits and knowledge sharing of cross-task and cross-modal learning, we conduct an analysis by selectively removing data from specific task categories during training. We design three variants: (1) OneThinker-wo-spatial-grounding, which excludes spatial grounding data; (2) OneThinker-wo-temporal-grounding, which removes all temporal grounding data; (3) OneThinker-wo-ImageQA, which omits all image QA samples.

As shown in table 10, removing either spatial or temporal grounding leads to a noticeable drop in performance across other tasks. In particular, the absence of temporal grounding significantly degrades results on video QA and tracking, indicating that temporal grounding may enhance the model's temporal perception and sequential reasoning ability. Similarly, removing spatial grounding results in lower accuracy on both image QA and segmentation, indicating that spatial localization tasks may contribute valuable structural and positional cues that benefit broader visual reasoning.

Moreover, excluding ImageQA causes the severe performance drop on video QA. We attribute this to the generally higher quality and greater diversity of image QA datasets, which help the model develop stronger general reasoning and recognition capabilities that transfer well to video understanding. This observation confirms that knowledge learned from static images can generalize to dynamic video scenarios, reflecting the benefit of cross-modal transfer.

Overall, these results suggest that certain tasks and modalities can benefit from others during joint training in OneThinker. By jointly training on diverse visual tasks, OneThinker effectively shares knowledge across domains and emerges as a multimodal reasoning generalist.

4.5 Zero-shot Generalization to Unseen Tasks

We further evaluate OneThinker's zero-shot generalization on unseen visual understanding tasks. These unseen tasks are selected from MMT-Bench [51], which contains 162 diverse visual tasks. As shown in fig. 5, OneThinker-8B clearly outperforms Qwen3-VL-Instruct-8B across multiple unseen tasks, such as point tracking, image quality assessment, GUI tasks and rotated obejct detection. These results demonstrate that unified multimodal reasoning enables the model to generalize beyond its training tasks, showing promising transferability to novel real-world scenarios.

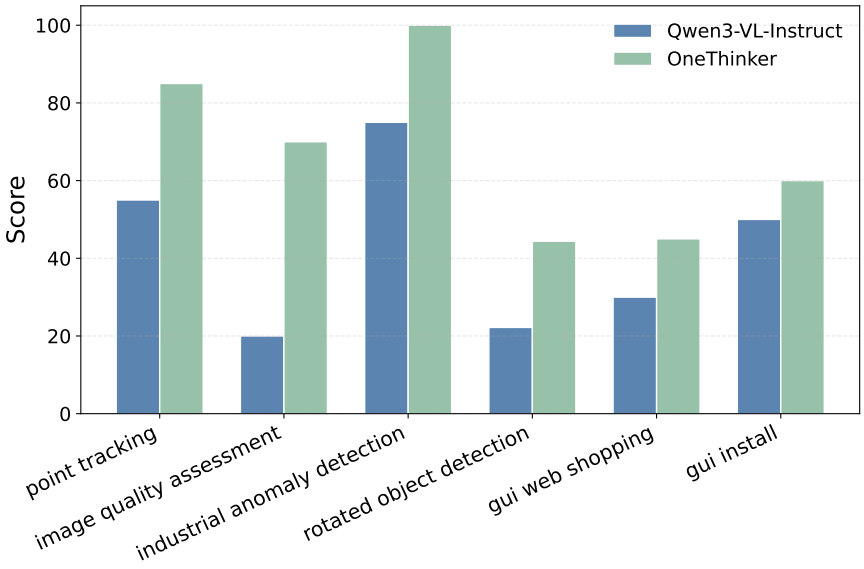


Figure 5: Performance on unseen visual tasks.

5 Conclusion

In this work, we present OneThinker, an all-in-one multimodal reasoning model that unifies diverse visual foundation tasks for images and videos. To support training, we construct OneThinker-600k dataset for RL training and its CoT-annotated subset OneThinker-SFT-340k for SFT cold start. We further propose EMA-GRPO, an RL algorithm that balances optimization across heterogeneous visual tasks through task-wise adaptive reward normalization. Extensive experiments demonstrate that OneThinker achieves strong performance across tasks. We hope this work takes a step toward scalable and unified multimodal reasoning generalist.

References

[1] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

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

[3] 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.

[4] 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.

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

[6] 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.

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

[8] Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *arXiv preprint arXiv:2505.15804*, 2025.

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

[10] 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.

[11] 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.

[12] Zuyao You and Zuxuan Wu. Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. *arXiv preprint arXiv:2506.22624*, 2025.

[13] 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.

[14] 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.

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

[16] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[17] 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.

[18] Michael Bereket and Jure Leskovec. Uncalibrated reasoning: Grpo induces overconfidence for stochastic outcomes. *arXiv preprint arXiv:2508.11800*, 2025.

[19] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025.

[20] Wenke Huang, Quan Zhang, Yiyang Fang, Jian Liang, Xuankun Rong, Huanjin Yao, Guancheng Wan, Ke Liang, Wenwen He, Mingjun Li, et al. Mapo: Mixed advantage policy optimization. *arXiv preprint arXiv:2509.18849*, 2025.

[21] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[22] 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.

[23] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019.

[24] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024.

[25] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[26] Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025.

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

[28] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, 2025.

[29] 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.

[30] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*, 2025.

[31] Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifu Luo, Yifei Zhao, Kai Qin, Xufei Lv, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement fine-tuning powers reasoning capability of multimodal large language models. *arXiv preprint arXiv:2505.18536*, 2025.

[32] Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, et al. Spacevista: All-scale visual spatial reasoning from mm to km. *arXiv preprint arXiv:2510.09606*, 2025.

[33] Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, et al. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. *arXiv preprint arXiv:2510.11718*, 2025.

[34] Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025.

[35] Shuang Chen, Yue Guo, Yimeng Ye, Shijue Huang, Wenbo Hu, Haoxi Li, Manyuan Zhang, Jiayu Chen, Song Guo, and Nanyun Peng. Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping. *arXiv preprint arXiv:2510.08457*, 2025.

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

[37] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025.

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

[39] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 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, 2025.

[40] Shihan Dou, Shichun Liu, Yuming Yang, Yicheng Zou, Yunhua Zhou, Shuhao Xing, Chenhao Huang, Qiming Ge, Demin Song, Haijun Lv, et al. Pre-trained policy discriminators are general reward models. *arXiv preprint arXiv:2507.05197*, 2025.

[41] Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language model. *arXiv preprint arXiv:2508.06206*, 2025.

[42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[43] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[44] 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.

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

[46] 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.

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

[48] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

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

[50] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016.

[51] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

[52] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[53] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[54] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

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

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

[57] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025.

[58] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 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, 2025.

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

[60] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[61] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025.

[62] Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025.

[63] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[64] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

[65] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

[66] 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.

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

[68] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[69] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[70] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024.

[71] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024.

[72] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

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

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

[75] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016.

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

[77] Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8592–8603, 2025.

[78] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678, 2024.

[79] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024.

[80] Biao Wang, Wenwen Li, and Jiawei Ge. R1-track: Direct application of mllms to visual object tracking via reinforcement learning. *arXiv preprint arXiv:2506.21980*, 2025.

[81] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2694–2703, 2023.

[82] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[83] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

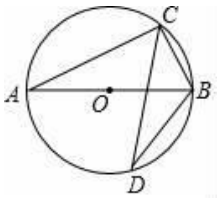
[84] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115. Springer, 2024.

[85] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022.

A Reasoning Examples

Image QA

Question: As shown in the figure, AB is the diameter of ⊙O, and points C and D are on ⊙O. If ∠ABD = 50.0, then the degree of ∠BCD is () Choices: (A) 30° (B) 35° (C) 40° (D) 45°



OneThinker-8B: <think>Got it, let's try to solve this problem step by step. First, let's recall some circle theorems. \n\nWe know that AB is the diameter of circle O. So, any angle inscribed in a semicircle is a right angle. That means ∠ACB should be 90° because triangle ACB is inscribed in the semicircle with AB as the diameter. Now, let's look at ∠ABD. The problem says ∠ABD is 50°. Let's think about quadrilateral or triangles here. Maybe we can use the inscribed angle theorem, which states that an inscribed angle is half the measure of its intercepted arc. First, let's consider arc AD or arc AB. Since AB is the diameter, the arc AB is 180°. ∠ABD is an inscribed angle that intercepts arc AD. So ∠ABD intercepts arc AD, so the measure of arc AD would be 2 × ∠ABD = 2 × 50° = 100°? Wait, maybe not. Let's check another angle. Alternatively, ∠BAD and ∠BCD might be related. Let's see. ∠BAD is an inscribed angle intercepting arc BD. ∠BCD is also an inscribed angle intercepting arc BD. So they might be equal or supplementary?

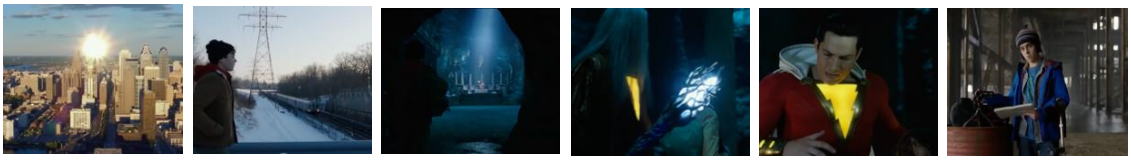
Wait, AB is the diameter, so ∠ADB is also a right angle (since angle inscribed in a semicircle). So triangle ADB is a right-angled triangle at D. In triangle ADB, ∠ABD is 50°, so ∠BAD = 90° - 50° = 40°. \n\nNow, ∠BCD and ∠BAD are inscribed angles that intercept the same arc BD. Because ∠BAD intercepts arc BD, and ∠BCD also intercepts arc BD. So they should be equal. Therefore, ∠BCD = ∠BAD = 40°. Let's verify again.\n\nAB is diameter, so ∠ACB = 90°, ∠ADB = 90°. In △ABD, ∠BAD = 90° - 50° = 40°. ∠BCD and ∠BAD both subtend arc BD, so they are equal. Hence, ∠BCD is 40°, which is option C.</think>

<answer>C</answer>

Figure 6: Reasoning example of image question answering task.

Video QA

Question: In the context of the scenes described, what is the most likely narrative purpose of the glowing key (0:00:20-0:00:30) and its connection to the \“SUPER STRENGTH TEST 1\” sign (0:00:50-0:01:00), given the recurring appearances of superheroes and promotional content for a film platform? (A) The key unlocks a hidden villain base (B) The key is a mystical artifact that activates Shazam’s powers, and the test sign marks a public demonstration of his abilities to gain public trust (C) The key grants access to a magical realm where the superhero team must prove their worth via strength tests (D) The key is a McGuffin distracting from the \"Don’t scream!\" scene, while the test sign is unrelated promotional material for Filmisnow trailers



OneThinker-8B: <think>So, let's try to figure out this question. The options are about the glowing key and the \"SUPER STRENGTH TEST 1\" sign in the context of a superhero film, likely Shazam! since there's a mention of Shazam and the red suit superhero. First, let's recall the elements: the glowing key, the test sign, and the superhero (Shazam) context. Shazam's story involves a magical artifact (the Shazam stone) that grants powers, and tests of strength.....Key unlocks villain base, test is training. But the context is promotional for a film, so maybe not villain base..... C. Key grants access to magical realm for strength tests. The snowy industrial area with the test sign—maybe a test location. But the key's connection? D. McGuffin distracting, test sign unrelated. Unlikely, since it's promotional.

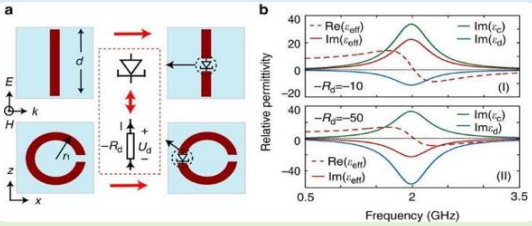
Wait, in Shazam, the Shazam stone is the key (glowing, magical). The \"SUPER STRENGTH TEST\" could be a test of his powers, maybe a public demonstration. But option C says \"magical realm where team proves worth\". Alternatively, the key is the Shazam stone, and the test is a public test. But let's think again.\n\nThe video is a trailer for Shazam, so the glowing key might be the Shazam stone, and the test sign is a test of his strength. But option C: \"key grants access to magical realm where superhero team proves worth via strength tests\". Maybe the key is a portal or artifact, and the test is part of that. Alternatively, the key is the stone, and the test is a public demo.\n\nWait, the correct answer is likely C..... The key (maybe the Shazam stone) is a magical artifact, and the test is a strength test in that realm.\n\nAlternatively, B: key activates Shazam’s powers, test sign public demo. But \"public trust\" might not be the main narrative purpose.\n\nAnother angle: Shazam's powers are activated by the stone, and tests of strength are part of his journey. The key (stone) is the source, and the test is a demonstration. But option C's \"magical realm\" might be the setting for the test.\n\nYes, C seems plausible. The key (maybe the stone) is a magical artifact, and the test is a strength test in a magical realm.</think>

<answer>C</answer>

Figure 7: Reasoning example of video question answering task.

Image Caption

Question: Give a description of this image.



OneThinker-8B: <think>So, let\'s try to describe this image. First, it\'s a scientific figure with two main parts, labeled (a) and (b)..... There are multiple curves: solid and dashed lines for real (Re) and imaginary (Im) parts of effective permittivity (ε_eff) and maybe another permittivity (ε_c, ε_d).....showing structural diagrams (a) and their electromagnetic properties (relative permittivity as a function of frequency) in (b), with parameters like Rd affecting the results.</think>

<answer>This image appears to be a scientific illustration, likely from a research paper in the field of physics or engineering, focusing on the study of metamaterials or resonators, particularly in the context of electromagnetic properties. The image is divided into two main sections, labeled (a) and (b).\n\n**Section (a):**\n- The left side shows two diagrams, one rectangular and one circular, both with a red-colored material, possibly representing a dielectric or resonant structure. The rectangular diagram has arrows indicating a dimension \'d\' and a wave vector \'k\' with an electric field \'E\' and magnetic field \'H\' direction, suggesting an electromagnetic wave interaction.\n- The right side of section (a) shows a schematic representation of an equivalent circuit model for these structures, with symbols indicating resistors and voltage sources, possibly representing the electrical behavior of the materials under study. There are dashed arrows indicating a transformation or correspondence between the physical structures and their circuit models.\n\n**Section (b):**\n- This section contains two graphs, labeled (I) and (II), plotting the relative permittivity against frequency in GHz.\n- Both graphs display curves for the real (Re) and imaginary (Im) parts of the effective permittivity (ε_eff) and another permittivity (ε_c), with different colors representing different values of a parameter, possibly a resistance \'R_d\' (with values -10 and -50 shown).\n- Graph (I) shows a higher variation in permittivity with frequency compared to graph (II), where the permittivity values are closer to zero, indicating a different response or material property at these specific resistance values.\n- The x-axis in both graphs is labeled "Frequency (GHz)" and the y-axis is "Relative permittivity," providing a quantitative analysis of how the permittivity of these materials changes with frequency under different conditions.\n\nOverall, the image seems to illustrate both the physical and circuit models of resonant structures and their corresponding electromagnetic properties, specifically focusing on how permittivity varies with frequency under different parameters, which is crucial for understanding and designing metamaterials or resonators for various applications in telecommunications, optics, or other fields requiring precise control over electromagnetic properties.</answer>

Figure 8: Reasoning example of image caption task.

Video Caption

Question: Give a description of this video.

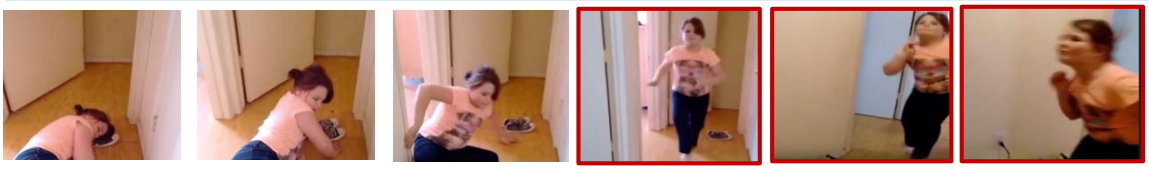
OneThinker-8B: <think>In a cozy, well-lit living room filled with bookshelves, a lamp, and a colorful couch, two individuals engage in an animated discussion. One person, seated on the couch, is focused on writing or taking notes in a notebook, wearing a vibrant, floral-patterned cardigan over a pink top. The other person.....Towards the end, the seated person appears to take notes or reflect, while the standing individual continues to speak, maintaining the educational or explanatory tone of the scene. The overall setting and interaction convey a friendly, intellectual exchange in a comfortable, lived-in space.</think>

<answer>The video takes place in a cozy, well-lit living room that exudes a warm, intellectual ambiance, characterized by a richly decorated space filled with bookshelves, a colorful striped couch, and a large whiteboard covered in scientific equations and diagrams. The room is bathed in a homely glow, enhanced by a lamp and a festive, beaded tree in the background, creating a setting that feels both scholarly and inviting. Two individuals are the focal points of the scene. One person, seated comfortably on the couch, is engrossed in writing or taking notes in a notebook, wearing a vibrant, floral-patterned cardigan over a pink top. Their focused demeanor suggests they are deeply engaged in the discussion or study session. The other person, standing near the whiteboard, is dressed in a blue shirt with colorful vertical stripes and yellow sleeves, and appears to be explaining or teaching, using a marker to point at and write on the whiteboard. Their animated gestures and active engagement indicate they are leading the discussion, possibly explaining complex scientific concepts. As the video progresses, the standing individual continues to elaborate on the whiteboard, gesturing towards various equations and diagrams with increasing emphasis, indicating a detailed and possibly intricate explanation. The seated person remains attentive, occasionally glancing up from their notebook with expressions of surprise or interest, reflecting their active participation and engagement in the session. The interaction is dynamic and educational, with the standing person frequently pointing and writing on the whiteboard, while the seated individual responds with focused listening and note-taking. The camera alternates between close-up shots of the individuals, capturing their expressions and actions in detail, and wider shots that provide a comprehensive view of the room and their interaction. Towards the end of the video, the seated person appears to be taking notes or reflecting, with a thoughtful expression, while the standing individual continues to explain, occasionally pausing and looking down, possibly to gather their thoughts. The overall atmosphere remains one of a lively, intellectual exchange, with both individuals deeply involved in a collaborative and educational session in a warm, inviting living room setting.</answer>

Figure 9: Reasoning example of video caption task.

Temporal Grounding

Question: Give the query: 'person runs back.', when does the described content occur in the video?



OneThinker-8B: <think>So, let's analyze the video. The query is 'person runs back.' Let's check each frame. Looking at the timestamps: The person was lying down, then gets up. Then, towards the end, the person runs towards the camera and then maybe back? Wait, the last part: the person runs towards the camera (maybe back direction). Let's see the movement. From the frames, around 15-22 seconds, the person gets up, walks, then runs. The action of running back—when does that happen? Let's check: The person starts lying on floor, then moves. Then, when they run towards the camera and then maybe back? Wait, the video shows the person running towards the camera (maybe towards the end). Let's check the time span.

Looking at the sequence: The person wakes up, moves, then runs. The running action towards the camera and then maybe back? The last part where the person runs towards the camera and then exits. The time when the person runs back—probably around 15-22 seconds. Let's confirm: The video shows the person lying down, then getting up, moving, then running. The action of running back would be when they run towards the camera and then maybe back. The correct time span for 'person runs back'—let's see the movement.

Yes, the person runs towards the camera (maybe back direction) in the later part. So the time span is around 15 to 22 seconds. Let's check:\n\n15-22 seconds: The person is moving quickly, running. So the time when the person runs back is in that range.</think>

<answer>{"time": [15.0, 22.0]}</answer>

Figure 10: Reasoning example of temporal grounding task.

Spatial Grounding

Question: Please provide the bounding box coordinate of the region this sentence describes: a bush of plant behind middle woman.

OneThinker-8B: <think>To find the bush behind the middle woman, first identify the middle woman: she is the one in the center, wearing a black shirt. Behind her is a large green bush. This bush sits in the background, positioned behind her as she is seated.</think>

<answer>{"boxes": [442, 56, 664, 357]}</answer>

Figure 11: Reasoning example of spatial grounding task.

Spatial-Temporal Grounding

Question: At which time interval in the video can we see an adult wearing blue pants plays the guitar occurring? When and where does the described content occur in the video? Please firstly give the start and end time, spatial bounding box corresponding to each integer second.

OneThinker-8B: <think>To find when an adult in blue pants plays the guitar, observe the video: the adult in blue pants playing guitar appears during a segment where the camera shifts focus. This occurs from around 5 seconds to 19 seconds. During this interval, the adult is seated on the right side of the frame, engaged in playing the guitar.</think>

<answer>{"time": [5.015, 19.511], "boxes": {"6": [670, 277, 998, 998], "7": [647, 281, 998, 998], "8": [616, 310, 998, 998], "9": [609, 327, 998, 998], "10": [609, 333, 998, 998], "11": [606, 333, 909, 998], "12": [600, 333, 889, 998], "13": [589, 333, 883, 998], "14": [583, 333, 877, 998], "15": [580, 333, 877, 998], "16": [577, 333, 877, 998], "17": [577, 333, 877, 998], "18": [577, 333, 877, 998], "19": [577, 333, 877, 998]}}</answer>

Figure 12: Reasoning example of spatial-temporal grounding task.

Tracking

Question: Given the bounding box [537,403,768,703] of the target object in the first frame, track this object in each frame and output its bounding box once per second.

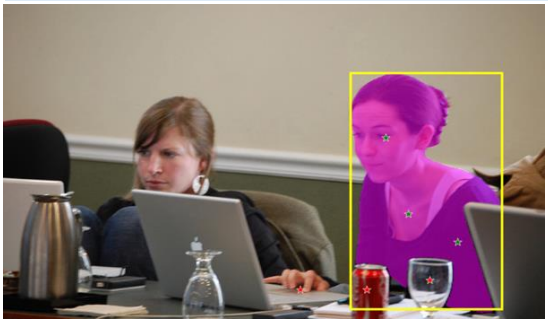
OneThinker-8B: <think>To track the white Ford Mustang convertible across each second up to 32 seconds, observe its movement pattern: the car travels forward along the road, so its position in the frame shifts subtly over time. Initially, the car occupies a central area of the frame. As seconds pass, the car moves closer to the camera, causing its bounding box to adjust incrementally. Over each second, the car's position shifts slightly toward the lower portion of the frame (as it approaches the camera) and its bounding box expands a bit to account for the closer perspective. Throughout the sequence, the car remains the central focus, with smooth, gradual changes in its placement and the size of the bounding box reflecting its forward motion and proximity to the camera.</think>

<answer>{"boxes": {"1": [537, 403, 768, 703], "2": [537, 353, 797, 690], "3": [532, 323, 827, 687], "4": [519, 282, 850, 690], "5": [515, 260, 877, 694], "6": [510, 246, 887, 694], "7": [509, 246, 897, 703], "8": [509, 232, 947, 710], "9": [509, 210, 972, 732], "10": [482, 200, 989, 787], "11": [469, 196, 995, 856], "12": [459, 203, 995, 929], "13": [422, 221, 995, 987], "14": [397, 244, 995, 994], "15": [389, 250, 995, 994], "16": [382, 250, 995, 994], "17": [370, 250, 995, 994], "18": [364, 250, 995, 994], "19": [362, 250, 995, 994], "20": [354, 244, 995, 994], "21": [350, 239, 995, 994], "22": [347, 239, 995, 994], "23": [347, 244, 965, 979], "24": [350, 244, 945, 944], "25": [350, 244, 904, 910], "26": [350, 244, 845, 860], "27": [347, 244, 797, 779], "28": [347, 244, 772, 750], "29": [350, 250, 764, 742], "30": [350, 250, 757, 739], "31": [350, 250, 750, 739], "32": [350, 250, 750, 739]}}</answer>

Figure 13: Reasoning example of tracking task.

Image Segmentation

Question: Given the referring expression "a woman touching a laptop wearing a white tank top under a black shirt", provide one bounding box, along with positive points located inside the object and negative points located outside, as segmentation hints.



OneThinker-8B: <think>Got it, let's tackle this task. First, identify the target object: \"a woman touching a laptop wearing a white tank top under a black shirt\". Let's look at the two women. The right woman has a black shirt with a white tank top underneath, and she's touching a laptop. So that's the target.\n\nNow the bounding box. Let's estimate coordinates. Let's see, the woman on the right. Let's set x1, y1, x2, y2. Let's check the image. She's on the right, so box around her. Let's approximate: x1 around 640, y1 around 278, x2 around 914, y2 around 957. Need to adjust.....The box around her would include her body, but maybe the table? Wait, she's sitting, so inside the box (her bounding box) might have parts like the table? No, the box is around her. Alternatively, if inside the box, non-object areas. But she's a person, so maybe the laptop she's touching? No, the laptop is part of the scene but she's the target. Alternatively, inside the box but not on her. Let's see.\n\nSo boxes: Let's define the box for the right woman. Let's say [642, 276, 917, 956] (adjusting). Then positive points: three points on her body. Like (700,450) [face], (754,650) [chest], (820,750) [arm]. Negative points: inside the box but not on her. Maybe on the laptop she's using? But the laptop is in front. Alternatively, the table items inside the box? Maybe the can or glass? .....Alternatively, adjust.\n\nLet's finalize:\n\nBox: Let's set x1=640, y1=275, x2=915, y2=955.\n\nPositive points: three points on her body. Like (704,460) (face), (745,680) (chest area), (834,760) (arm).\n\nNegative points: inside the box but not on her. Maybe on the laptop she's using? But the laptop is in front. Alternatively, the table items inside the box. For example, the red can, glass, etc. But need to be off the woman. So points on the table inside the box but not on her. Let's pick (670,900) (red can), (784,870) (glass), (550,900) (maybe a mouse? But need to check). Alternatively, inside the box but on background. Let's confirm.</think>

<answer>{"boxes": [640, 275, 915, 955], "positive_points": [[704, 460], [745, 680], [834, 760]], "negative_points": [[670, 900], [784, 870], [550, 900]]}</answer>

Figure 14: Reasoning example for an image segmentation task. The resulting answer will be forwarded to SAM2 to produce the mask.

Video Segmentation

Question: Given the referring expression “Which creature in the footage poses the greatest threat to an unwanted visitor encroaching on its domain?”, Provide the key frame time along with one bounding box, as well as positive points inside the object and negative points outside, to serve as segmentation hints.

OneThinker-8B: <think>So, let's tackle this task. The question is about identifying the creature that poses the greatest threat to an unwanted visitor in its domain. In the video, elephants are the main subjects, and the largest elephant would be the most dominant, so that's the target. Let's pick a time frame. Let's choose 1.0 second as a representative time.\n\nFirst, the bounding box for the large elephant. Let's look at the frame around 1.0s. The big elephant is the main one in the foreground. Let's estimate coordinates. Let's say the box is around the large elephant. Let's set x1, y1, x2, y2. Let's approximate: x1 around 180, y1 around 15, x2 around 826, y2 around 940 (adjusting for the video's aspect).

Positive points: inside the elephant. So pick three points on the body, like torso, head, leg.\n\nNegative points: need to be outside the object. If inside the box, maybe background. But the elephant is solid, so maybe inside the box but on the ground? Wait, the elephant's body is the object. So negatives could be on the ground inside the box but not on the elephant. Or outside.

Let's confirm: Time chosen: 1.0 seconds. Box around the large elephant. Let's set boxes as [175, 10, 828, 945] (adjusting).\n\nPositive points: three points inside the elephant. For example, (350, 400), (500, 300), (650, 500) — these are on the body.\n\nNegative points: three points outside. Maybe on the ground inside the box but not on the elephant? Or outside. Let's pick inside the box but on the ground. Like (200, 850) (on the ground near the elephant's leg), (750, 850) (near the leg), and (400, 900) (ground). These are outside the elephant's body.

Wait, need to ensure they are strictly off the object. So yes, on the ground inside the box but not on the elephant. That works.</think>

<answer>{"time": 1.0, "boxes": [175, 10, 828, 945], "positive_points": [[350, 400], [500, 300], [650, 500]], "negative_points": [[200, 850], [750, 850], [400, 900]]}</answer>

Figure 15: Reasoning example for an video segmentation task. The resulting answer will be forwarded to SAM2 to produce the mask.

B Prompt Template

System Prompt

Please answer this question based on the visual content. Provide your thinking process between the <think> and </think> tags, and then give your final answer between the <answer> and </answer> tags. At the end, you must output the final answer in the format: <answer><your_answer_here></answer>

Figure 16: System prompt for all tasks.

Prompt for QA

```
"multiple choice": (
"Please provide only the single option letter (e.g., A, B, C, D, etc.) "
"within the <answer>...</answer> tags.\n"
"Example:\n<answer>A</answer>"
),

"numerical": (
"Please provide only the numerical value within the <answer>...</answer> tags.\n"
"Example:\n<answer>3.14</answer>"
),

"OCR": (
"Please provide only the transcribed text within the <answer>...</answer> tags.\n"
"Example:\n<answer>Hello World</answer>"
),

"open-ended": (
"Please provide only your text answer within the <answer>...</answer> tags.\n"
"Example:\n<answer>The capital of France is Paris.</answer>"
),

"regression": (
"Please provide only the numerical value within the <answer>...</answer> tags.\n"
"Example:\n<answer>42.7</answer>"
),

"math": (
"Please provide only the final result (a number or LaTeX formula) within the
<answer>...</answer> tags.\n"
"Example:\n<answer>$$-\\dfrac{3}{2}$$</answer>"
)
```

Figure 17: Prompt for QA tasks.

Prompt for Grounding and Tracking

```
"temporal grounding": (
"Please provide only the time span in seconds as JSON within the <answer>...</answer> tags. Example:\n<answer>{\"time\": [12.3, 25.7]}</answer>"
),

"spatial grounding": (
"Please provide only the bounding box as JSON with key 'boxes' within the <answer>...</answer> tags.\n"
"Example:\n<answer>{\"boxes\": [35, 227, 437, 932]}</answer>"
),

"spatial-temporal grounding": (
"Please provide only the time span in seconds and bounding boxes as JSON within the <answer>...</answer> tags. You MUST output one bounding box for every integer second within the given time span (inclusive).\n"
"Example: <answer>{\"time\": [8.125, 13.483], \"boxes\": {\"9\": [317, 422, 582, 997], "
"\"10\": [332, 175, 442, 369], \"11\": [340, 180, 450, 370]}}</answer>\n"
"Note: Each key in 'boxes' must be an integer second within the span, and its value must be a 4-number bounding box [x1, y1, x2, y2]."
),

"tracking": (
"Please track the target object throughout the video and provide one bounding box per second, within the <answer>...</answer> tags.\n"
"Example:"<answer>{\"boxes\": {\"1\": [405, 230, 654, 463], \"2\": [435, 223, 678, 446], ..., \"32\": [415, 203, 691, 487]}}</answer>\n"
)
```

Figure 18: Prompt for grounding and tracking tasks.

Prompt for Segmentation

```
"image_segmentation": (
"This task prepares inputs for image object segmentation with a specialized model (e.g., SAM2).\n"
"Please provide ONE bounding box, 3 positive points (clearly INSIDE the object), and 3 negative points (clearly OUTSIDE the object) within the <answer>...</answer> tags.\n"
"Choose informative points that help distinguish object vs. background. Prefer negatives on clear non-object pixels INSIDE the box when safe; otherwise place them just outside on obvious background. Negatives must NEVER be on the object or on its boundary.\n"
"Example:\<answer>{\"boxes\": [x1, y1, x2, y2], \"positive_points\": [[x,y],[x,y],[x,y]], \"negative_points\": [[x,y],[x,y],[x,y]]}</answer>"
),

"video_segmentation": (
"This task prepares inputs for video object segmentation with a specialized model (e.g., SAM2).\n"
"Please select ONE representative time (in seconds), and provide ONE bounding box, "
"3 positive points (clearly INSIDE the object), and 3 negative points (clearly OUTSIDE the object) within the <answer>...</answer> tags.\n"
"Choose informative points that help distinguish object vs. background. Prefer negatives on clear non-object pixels INSIDE the box when safe; otherwise place them just outside on obvious background. Negatives must NEVER be on the object or on its boundary.\n"
"Example:\<answer>{\"time\": <time_in_seconds>, \"boxes\": [x1, y1, x2, y2], "
"\"positive_points\": [[x,y],[x,y],[x,y]], \"negative_points\":[[x,y],[x,y],[x,y]]}</answer>"
)
```

Figure 19: Prompt for segmentation tasks.