Mimic Human Cognition, Master Multi-Image Reasoning: A Meta-Action Framework for Enhanced Visual Understanding

Supplementary Material

7. Background: DAPO

DAPO [93] is an improved variant of GRPO [19], which directly computes the advantage A_t using the average reward over multiple sampled outputs, thereby eliminating the need for a separate value function as in PPO. Specifically, given a prompt $\mathbf{q} \sim P(Q)$, we sample G rollouts $\{\mathbf{o}_i\}_{i=1}^{G}$ from the current policy $\pi_{\theta_{\text{old}}}$. At each token position t in rollout i, the likelihood ratio is defined in Eq. 3.

$$r_{i,t}(\theta) = \frac{\pi_\theta\big(o_{i,t} \mid \mathbf{q},\, \mathbf{o}_{i,<t}\big)}{\pi_{\theta_{\text{old}}}\big(o_{i,t} \mid \mathbf{q},\, \mathbf{o}_{i,<t}\big)} \tag{3}$$

The group-relative advantage $\hat{A}_{i,t}$ is then obtained by standardizing each return R_i within the group, defined in Eq. 4.

$$\hat{A}_{i,t} = \frac{R_i \;-\; \text{Mean}\big(\{R_j\}_{j=1}^{G}\big)}{\text{Std}\big(\{R_j\}_{j=1}^{G}\big)}. \tag{4}$$

In contrast to GRPO, DAPO introduces several methodological advancements. Specifically, it employs a Clip-Higher mechanism, wherein ϵ_{high} is set greater than ϵ_{low} to enhance exploratory behavior; integrates Dynamic Sampling to systematically exclude data instances lacking informative learning signals; incorporates an Overlong Punishment strategy to constrain excessively verbose outputs; and adopts a Token-level Loss formulation to mitigate the inherent bias between responses of varying lengths. The training then proceeds by maximizing the clipped surrogate objective, defined for DAPO as follows:

$$\begin{aligned}
&\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\
&\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\right. \\
&\qquad\left.\text{clip}\big(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\big)\hat{A}_{i,t}\Big)\right], \\
&\text{s.t.}\quad 0 < \Big|\{o_i \mid \text{is_equivalent}(R_i, 1)\}\Big| < G.
\end{aligned} \tag{5}$$

8. Benchmark

This section provides a detailed description of the benchmark used for evaluation.

MUIR MUIRBENCH [71] is a comprehensive benchmark designed for robustly evaluating MLLMs' multi-image understanding capabilities. It comprises 11,264 images and 2,600 multiple-choice questions (average 4.3 images per instance), covering 12 diverse multi-image tasks (e.g., action understanding, cartoon storytelling, geographic map reasoning, 3D object multiview retrieval).

MMIU The Multimodal Multi-image Understanding (MMIU) [58] is a comprehensive benchmark tailored for evaluating MLLMs on multi-image comprehension tasks. Structured around cognitive psychology, it enumerates 7 types of multi-image relationships (refined from semantic, temporal, spatial categories) and covers 52 diverse tasks (e.g., multi-view action recognition, 3D object detection) . In terms of scale, MMIU includes 77,659 images (2–32 per instance, averaging 6.64) and 11,698 meticulously curated multiple-choice questions.

MV-MATH MV-MATH [74] is a specialized benchmark designed to evaluate MLLMs on mathematical reasoning in multi-visual contexts—addressing the gap in existing benchmarks that mostly focus on single images. It comprises 2,009 high-quality mathematical problems derived from real K-12 scenarios.

EMMA EMMA [20] is a benchmark designed to evaluate Multimodal LLMs on genuine cross-modal reasoning. Its 2,788 questions across math, physics, chemistry, and coding require integrated visual-textual understanding, preventing solutions based on shallow cues or text alone.

Mantis-Eval Mantis-Eval [32] is a benchmark dataset designed to evaluate a model's ability to reason across multiple images. It contains 217 challenging examples.

MIRB MIRB [98] is a dedicated dataset addressing the gap in evaluating vision-language models (VLMs) on multi-image understanding, as existing benchmarks focus primarily on single-image inputs. It encompasses 925 samples across four core dimensions: perception, visual world knowledge, reasoning, and multi-hop reasoning, with all tasks requiring cross-comparison of multiple images (ranging from 2 to 42, averaging 3.78 per question).

MVBench MVBench [41] is a multi-modal video benchmark addressing the lack of temporal understanding evalu-

ation in MLLMs, covering 20 multi-frame-dependent video tasks (defined via a static-to-dynamic method). It is built efficiently by auto-converting public video annotations into multiple-choice QA (with ground-truth for fairness), reveals existing MLLMs' poor temporal understanding.

Video-MME Video-MME [18] is the first comprehensive benchmark designed to evaluate MLLMs in video analysis. It fills the gap in assessing the understanding of sequential visual data by featuring 900 videos (ranging from 11 seconds to 1 hour) across 6 core domains (e.g., Knowledge, Sports Competition) and 30 subfields. Each video is paired with three expert-annotated multiple-choice QA pairs, resulting in a total of 2,700 questions. To support multi-modal reasoning, the benchmark also provides subtitles for 744 videos and audio tracks for all 900 videos.

Video-MMMU Video-MMMU [24] is a benchmark designed to evaluate the knowledge acquisition capabilities of MLLMs from professional video content. It comprises 300 expert-level videos spanning six disciplines and 30 subfields, paired with 900 human-annotated question–answer pairs. The benchmark measures performance across three cognitive stages: (1) *Perception*, assessing whether models can extract salient knowledge-related details from video content; (2) *Comprehension*, evaluating the ability to grasp and reason about the underlying concepts; and (3) *Adaptation*, examining whether models can transfer the acquired knowledge to novel or unfamiliar scenarios.

MMMU-Pro MMMU-Pro [94] is an enhanced version of the MMMU benchmark, designed to more rigorously evaluate multimodal models' understanding and reasoning.

M3CoT M3CoT [9] addresses gaps in existing MCoT benchmarks (lack of visual reliance, single-step reasoning, limited domains) by enabling multi-domain, multi-step, multi-modal reasoning across 3 domains (science, mathematics, commonsense), 17 topics, and 263 categories. It has 11,459 total samples (7,973 train, 1,127 dev, 2,359 test) with diverse image types (geographic graphs, health images, etc.).

MM-MATH MM-MATH[68]consists of 5,929 open-ended middle school math problems paired with visual contexts, and it adopts fine-grained classification covering three dimensions: difficulty, grade level, and knowledge points.

MathVista MathVista [55] is proposed as a benchmark integrating challenges from mathematical and visual tasks. It contains 6,141 examples, sourced from 28 existing multimodal math datasets and 3 new ones (IQTest, FunctionQA, PaperQA), requiring fine-grained visual understanding and compositional reasoning—tasks that state-of-the-art foundation models find challenging.

MATH-V MATH-V [73] is a curated dataset designed to address the limited question diversity and subject breadth of existing visual math reasoning benchmarks. It comprises 3,040 high-quality math problems with visual contexts, all sourced from real math competitions.

9. Training Data Construction

For the construction of the training dataset, we referenced Mantis [32], LLaVA-Interleave [38], Leopard [31], and VideoR1 [17]. Overall, our dataset consists of multi-image data and single-image data, with 57k samples for cold-start training and 58k samples for RL. The detailed dataset statistics are presented in Table 1. Regarding the partitioning criteria for RL data and cold-start data, the key distinction between our cold-start and reinforcement learning dataset splits lies in the trajectory generation process described in Section 3.2. Cold-start training data consists of problems where GPT-4o successfully provides correct answers during Step 2, and for these instances, we proceed to Step 3 (retrieval-based diverse sampling) to obtain two distinct correct reasoning trajectories per problem that serve as supervised learning targets for cold start training. In contrast, reinforcement learning data comprises problems where GPT-4o fails to produce correct answers during Step 2, and these challenging cases are reserved for reinforcement learning.

10. Implementation

Our SFT experiments are primarily conducted using the LLaMA Factory framework [99], with the main hyperparameters summarized in Table 6. For the RL stage, we rely on the EasyR1 framework [90], a multi-model large-scale training system built upon VERL [64], and the key parameters are reported in Table 7. The experiments run on 32 A800 GPUs.

11. Case Study

Here we present a case study of our model in Figure 5 and 6, covering multi-image benchmarks, video benchmarks, and single-image benchmarks. The results demonstrate that, across different types of tasks, our model can dynamically invoke appropriate meta-actions to analyze the problem and produce correct answers. In Figure 7, we present multiple reasoning results for a single problem. It can be observed that the model explores different reasoning paths for the same problem, all of which lead to the correct answer.

Type	Dataset	Count for SFT	Count for RL
Multi-Image	ChartVQA[31]	2501	-
	SlideVQA[31]	3249	3000
	ALFRED[65]	8357	-
	Nuscenes[4]	-	4946
	RecipeQA[83]	8759	5069
	IconQA[53]	5315	3000
	nlvr2[67]	5424	1620
	Spot-the-Diff[30]	2248	2589
	LRV[47]	-	2993
	RAVEN[95]	-	3200
Video	Star[78]	5490	2754
	NextQA[79]	1193	3000
	Clevrer[92]	3047	4478
	Perception[61]	2964	2500
Single-Image	Clevr_cogen_a_train[1]	1506	-
	Clevr_CoGenT_TrainA_70K_Complex[2]	1159	3000
	M3COT[9]	1147	-
	Share-GRPO[89]	1145	3000
	GEOQA_R1V_Train_8K[3]	800	4816
	AI2D[35]	630	-
	MMK12[57]	442	3537
	Geometry3k[52]	317	1406
	ScienceQA[54]	259	-
	PISC[39]	244	-
	Geoqa+[5]	172	891
	GQA[27]	119	-
	CLEVR_v1.0[34]	118	-
	COCO[46]	78	-
	LRV[47]	-	3063

Table 5. Statistics of Training Data

[1] https://huggingface.co/datasets/leonardPKU/clevr_cogen_a_train

[2] https://huggingface.co/datasets/MMInstruction/Clevr_CoGenT_TrainA_70K_Complex

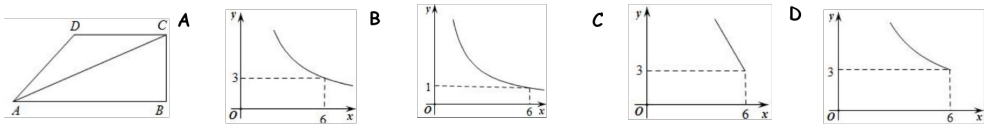
[3] https://huggingface.co/datasets/leonardPKU/GEOQA_R1V_Train_8K

Question:
What is the correct order of the images? A: [0, 1, 2, 3]\nB: [1, 3, 2, 0]\nC: [3, 2, 1, 0]\nD: [1, 3, 0, 2]

Response:
<**global**>Image2->cutting potatoes and onions->Image1, Image1->cooking the tomato-based mixture with chilies->Image4, Image4->adding tomatoes and seasoning to the pot->Image3</**global**>\n<**think**>The context involves preparing a dish step by step, and the images must match the sequence of actions described. The correct order must ensure that the cooking process aligns with the logical flow of actions described in the recipe. Observing the images and their relationships is key to matching them to the steps in the recipe context.</**think**>\n<**answer**>D</**answer**>

(a) Case1.

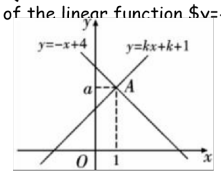
Question:
As shown in the image, in quadrilateral $A B C D$, $\\angle B=90^{\\circ}, A C=6, A B \\parallel C D$, and $A C$ bisects $\\angle D A B$. Let $A B=x$ and $A D=y$. The graph that can roughly represent the functional relationship between y and x is ( )



Response:
<**global**>Image1->analyze the geometry of quadrilateral $ABCD$->Image2</**global**> \n<**think**>The quadrilateral \\(ABCD\\) is given with \\(\\angle B = 90^\\circ\\), \\(AB \\parallel CD\\), and \\(AC\\) bisecting \\(\\angle DAB\\). This means we can use triangle similarity and trigonometric relationships to establish a functional relationship between \\(x\\) and \\(y\\). Specifically, we need to calculate \\(y\\) based on \\(x\\), and match the correct graph that represents this relationship. The key is to ensure that \\(y > 0\\) and \\(x > 0\\), and that \\(AC = 6\\).</**think**> \n<**hint**>The critical point is to correctly analyze the geometry and establish the relationship \\(y\\) in terms of \\(x\\). Since \\(AB \\parallel CD\\), we can use similar triangles to relate \\(x\\) and \\(y\\). The graph must satisfy the constraints \\(y > 0\\), \\(x > 0\\), and the functional relationship derived must match one of the options provided. Pay attention to the shape and position of the graph relative to the axes and key points.</**hint**> \n<**answer**>D</**answer**>

(b) Case2.

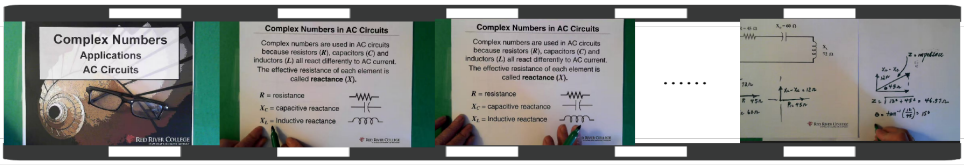
Question: As shown in the figure, it is known that the graph of the linear function $y=kx+k+1$ intersects the graph of the linear function $y=-x+4$ at point $A(1, a)$. What are the values of a and k?



Response:
<**focus**>The given images show two linear functions \\(y = -x + 4 \\) and \\(y = kx + k + 1 \\). The graphs intersect at point \\(A(1, a) \\). The task is to find \\(a \\) and \\(k \\).</**focus**>\n\n<**think**>1. Since \\(A(1, a) \\) lies on \\(y = -x + 4 \\), substitute \\(x = 1 \\) into \\(y = -x + 4 \\):\n\\[\ny = -1 + 4 = 3 \\quad \\Rightarrow \\quad a = 3.\n\\]\n2. The point \\(A(1, 3) \\) also lies on \\(y = kx + k + 1 \\). Substitute \\(x = 1 \\) and \\(y = 3 \\) into \\(y = kx + k + 1 \\):\n\\[\n3 = k(1) + k + 1 \\quad \\Rightarrow \\quad 3 = k + k + 1 \\quad \\Rightarrow \\quad 3 = 2k + 1 \\quad \\Rightarrow \\quad 2k = 2 \\quad \\Rightarrow \\quad k = 1.\n\\]\n</**think**>\n\n<**focus**>The solution steps confirm that \\(a = 3 \\) and \\(k = 1 \\). Verify consistency of calculations for correctness.</**focus**>\n\n<think>The calculations are consistent and logical. Ensure no arithmetic mistakes occurred during substitution and solving equations.</**think**>\n\n<**answer**>\\(a = 3 \\) and \\(k = 1 \\).</**answer**>

(c) Case3.

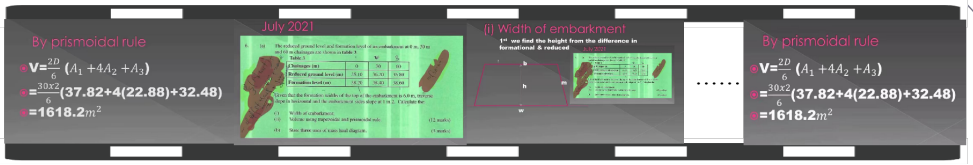
Figure 5. Case study.



Question:
In the AC circuit shown, the source voltage is _____(1) V at an angle of _____(2) degrees, and the difference between XL and XC is _____(3) \\Omega.Please select from the following choices.
A. 120, 45, 12 B. 60, 0, 12 C. 72, 0, 45 D. 45, 0, 72 E. 120, 0, 45 F. 60, 0, 72 G. 45, 0, 12 H. 72, 0, 60 I. 120, 0, 12 J. 120, 45, 72

Response:
<global>Image1->introduce the topic of AC circuits and complex numbers->Image2, Image2->explain the concept of reactance and its components->Image3, Image3->illustrate how different elements react to current and voltage->Image4, Image4->introduce the phasor representation of AC voltage and current->Image5, Image5->show the AC circuit diagram and component placement->Image6, Image6->highlight the values of R, XL, and XC->Image7, Image7->add the difference between XL and XC->Image8, Image8->begin to calculate the resultant phasor->Image9, Image9->continue calculating the resultant phasor and write intermediate steps->Image10, Image10->finalize the calculation of Z and Θ->Image14**</global>**\n\n**<focus>**The AC circuit diagram in the images shows a source voltage of 120 V at an angle of 0°, with R = 45 Ω, XC = 60 Ω, and XL = 72 Ω. The difference between XL and XC is 12 Ω (72 - 60 = 12).**</focus>**\n\n**<think>**The question asks for the source voltage, angle, and the difference in impedances. From the images, we can confirm that the source voltage is 120 V at 0° (AC 120 V ∠0°), and the difference between inductive reactance (XL) and capacitive reactance (XC) is 12 Ω (72 - 60 = 12). By matching the options, we can select the correct values that fit the description.**</think>**\n\n**<answer>**I**</answer>**

(a) Case4.



Question:
Calculate the side widths and cross-sectional area of an embankment (Figure 11.13) having the following dimensions:Road width = 20 m;existing ground slope = 1 in 10 (10%);Side slopes = 1 in 2 (50%);centre height = 10 m. Please select from the following choices.\nA. Side width w_{1} = 37.5 m,Side width w_{2} = 25.5 m,Total area = 420.75 m^{2}\nB. Side width w_{1} = 37.3 m,Side width w_{2} = 25.0 m,Total area = 418.75 m^{2}\nC. Side width w_{1} = 37.4 m,Side width w_{2} = 25.0 m,Total area = 418.75 m^{2}\nD. Side width w_{1} = 37.2 m,Side width w_{2} = 25.0 m,Total area = 418.75 m^{2}\nE. Side width w_{1} = 36.5 m,Side width w_{2} = 24.5 m,Total area = 408.75 m^{2}\nF. Side width w_{1} = 38.0 m,Side width w_{2} = 26.5 m,Total area = 425.75 m^{2}\nG. Side width w_{1} = 37.5 m,Side width w_{2} = 25.0 m,Total area = 410.75 m^{2}\nH. Side width w_{1} = 37.5 m,Side width w_{2} = 25.0 m,Total area = 418.75 m^{2}\nI. Side width w_{1} = 38.5 m,Side width w_{2} = 26.0 m,Total area = 428.75 m^{2}\nJ. Side width w_{1} = 37.0 m,Side width w_{2} = 24.0 m,Total area = 415.75 m^{2}

Response:
<global>Image7->introduce the formula for calculating width->Image8, Image8->provide values for height and side slope->Image9, Image9->calculate individual widths using formula->Image10, Image10->sum up to find total widths->Image12, Image12->introduce volume calculation using trapezoidal rule->Image13, Image13->perform volume calculation step by step->Image14**</global>**\n**<focus>**The key calculations involve determining the widths of the embankment using the formula \\(W = \\frac{1}{2}b + mh \\), where \\(b \\) is the formation width, \\(h \\) is the height, and \\(m \\) is the side slope. The total width is the average of the top and bottom widths. Additionally, the volume is calculated using the trapezoidal rule: \\(V = \\frac{D}{2} \\left((A_1 + A_n) + 2(A_2 + A_3 + ...) \\right) \\), where \\(A_1, A_2, A_3 \\) are areas at different levels, and \\(D \\) is the distance between levels.**</focus>**\n**<think>**1. Using the formula \\(W = \\frac{1}{2}b + mh \\), calculate the widths for different levels:\n - For \\(h = 3.1 \\): \\(W_1 = \\frac{1}{2}(6) + 3.1(1/2) = 3.0 + 1.55 = 37.5 \\, \\text{m} \\),\n - For \\(h = 2.2 \\): \\(W_2 = \\frac{1}{2}(6) + 2.2(1/2) = 3.0 + 1.1 = 25.0 \\, \\text{m} \\),\n - For \\(h = 2.8 \\): \\(W_3 = \\frac{1}{2}(6) + 2.8(1/2) = 3.0 + 1.4 = 25.0 \\, \\text{m} \\).\n \n2. Sum the widths to find the total widths:\n - \\(W_{\\text{total}} = 37.5 + 25.0 + 25.0 = 87.5 \\, \\text{m} \\).\n\n3. Use the trapezoidal rule to calculate volume:\n - \\(V = \\frac{30}{2} \\left(37.82 + 32.48 + 2(22.88) \\right) = 15 \\times (37.82 + 32.48 + 45.76) = 15 \\times 116.06 = 1740.9 \\, \\text{m}^3 \\).\n\n4. Match the calculated values to the options to find the correct one:\n - The correct option must match \\(W_{\\text{total}} = 87.5 \\, \\text{m} \\) and \\(V = 1740.9 \\, \\text{m}^3 \\). Option H satisfies these values with \\(W_1 = 37.5 \\, \\text{m}, W_2 = 25.0 \\, \\text{m}, \\text{and Total area} = 418.75 \\, \\text{m}^2 \\).**</think>**\n**<answer>**H**</answer>**

(b) Case5.

Figure 6. Case study.

Reasoning Path Analysis

Response: <think>The question involves identifying laboratory test results that are most likely to be associated with a 50-year-old woman who has had a high fever and papilledema, based on the provided images and options. Key points to consider include inflammation, infection, and conditions that might affect the ESR or other tests mentioned in the images. Papilledema and a high fever suggest systemic inflammation or infection, which may lead to elevated sedimentation rates or other markers of disease activity.</think>\n\n<global>Image1->introduces ESR and its measurement->Image5, Image5->explains the factors affecting sedimentation rates->Image9, Image9->highlights causes of elevated ESR->Image12</global>\n\n<think>The images emphasize that the ESR (erythrocyte sedimentation rate) is a key indicator of inflammation or infection. Since the question mentions a high fever and papilledema, these signs strongly suggest an inflammatory or infectious process. Among the options, an elevated ESR (50 mm/hr for a 50-year-old woman) aligns with the clinical scenario described. Other options (like complement levels, potassium, or specific enzymes) are less directly tied to systemic inflammation or papilledema.</think>\n\n<focus>The key image to focus on is Image12, which lists conditions affecting the ESR and highlights that inflammation and infection often elevate it, while anemia and certain other conditions can lower it. This supports the choice of an elevated ESR for this scenario.</focus>\n\n<hint>Error-prone points include confusing the significance of ESR with other laboratory values or overlooking that an elevated ESR is a common marker for inflammation or infection. It is crucial to match the clinical signs (fever, papilledema) with laboratory evidence (ESR increase).</hint>\n\n<answer>D</answer>

Trajectory: <think></think><global></global><think></think><focus></focus><hint></hint><answer></answer>

Response: <global>Image1->introduces ESR and sedimentation process->Image5, Image5->explains how to calculate and report ESR->Image6, Image8->normalizes ESR based on age and gender->Image12</global>\n<hint>The question requires identifying a laboratory test result that matches the given clinical scenario. Key points to consider are the patient's age (50 years), high fever, and papilledema, which suggest inflammation or systemic disease. The ESR is a measure of inflammation and sedimentation, and its normal range for women over 50 is 30 mm/hr or less, so an elevated ESR is likely in this case.</hint>\n<think>The presence of high fever and papilledema indicates a systemic inflammatory process or disease. Since ESR is a test that reflects inflammation, and the normal ESR for women over 50 is 30 mm/hr or less, an ESR of 50 mm/hr aligns with the clinical presentation of inflammation or disease activity. Other options (e.g., complement, creatinine, potassium, or specific enzymes) do not directly or clearly correlate with the provided symptoms or clinical picture in this case.</think>\n<focus>The key content to focus on is the relationship between inflammation, fever, and papilledema, and how these connect to laboratory tests. Specifically, ESR is mentioned as a measure of inflammation and its normal ranges by age and gender, as shown in the images and notes.</focus>\n<think>Given that the question asks for a likely laboratory test result and the images and text emphasize ESR as a marker for inflammation, and the normal ESR for women over 50 is 30 mm/hr or less, an ESR of 50 mm/hr is a plausible and significant result in this context, as it reflects the systemic inflammation associated with the symptoms described.</think>\n<answer>D</answer>

Trajectory: <global></global><hint></hint><think></think><focus></focus><think></think><answer></answer>

Figure 7. Reasoning Path Analysis.

Parameter	Value
Model	
model_name_or_path	Qwen2.5-VL-7B-Instruct
image_max_pixels	100352
Method	
stage	sft
do_train	true
finetuning_type	full
Dataset	
template	qwen2_vl
cutoff_len	12000
overwrite_cache	true
preprocessing_num_workers	16
dataloader_num_workers	4
Train	
per_device_train_batch_size	1
gradient_accumulation_steps	4
learning_rate	1.0e-5
num_train_epochs	2
lr_scheduler_type	cosine
warmup_ratio	0.1

Table 6. Hyperparameters used in SFT.

Parameter	Value
Data	
max_prompt_length	4096
max_response_length	4096
rollout_batch_size	64
max_pixels	100352
min_pixels	50176
Algorithm	
adv_estimator	grpo
kl_coef	0.0
filter_groups_enable	true
filter_max_num_gen_batches	20
filter_metric	acc
Worker.Actor	
global_batch_size	32
max_grad_norm	1.0
entropy_coeff	0.0
kl_loss_coef	0.0
clip_ratio_low	0.2
clip_ratio_high	0.28
optim.lr	1.0e-6
optim.weight_decay	1.0e-2
Worker.Rollout	
temperature	1.0
top_p	1.0
top_k	-1
n	8

Table 7. Hyperparameters used in RL.