

## A. RLVR for VLMs

Our model training method is based on the Group Relative Policy Optimization (GRPO) algorithm [29]. The core idea of GRPO is to sample multiple responses from the policy model for a given task and use verifiable rule-based rewards along with group relative advantages to replace the value model required in PPO, thereby reducing VRAM overhead.

Given a multimodal input  $Q = \{i, q\}$ , where  $i$  denotes the image input and  $q$  denotes the text question input; the policy model  $\pi_\theta$  outputs a set of responses  $\{o_1, \dots, o_G\}$ ; a rule-based reward function scores these responses, yielding a set of scalar rewards  $\{r_1, \dots, r_G\}$ . The GRPO algorithm normalizes the group-relative rewards to obtain the response-level advantage:

$$\hat{A}_i = \frac{r_i - \text{mean}(r_1, \dots, r_G)}{\text{std}(r_1, \dots, r_G) + \epsilon} \quad (7)$$

where  $\epsilon$  is a small value used for numerical stability.

Similar to PPO [28], GRPO also employs a clipped objective on the importance weights and uses a KL divergence mechanism to maintain the stability of the training updates, preventing excessive deviation from the old policy distribution. The loss function is as follows:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E} \left[ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left( \min \left( r_{i,t}(\theta) \hat{A}_i, \text{clip}(r_{i,t}(\theta), 1 - \epsilon_l, 1 + \epsilon_h) \hat{A}_i \right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right) \right] \quad (8)$$

where,

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} | q, o_{i,<t})}{\pi_{\text{old}}(o_{i,t} | q, o_{i,<t})}. \quad (9)$$

Here, we follow the work on DAPO [43] by decoupling the  $\epsilon$  values ( $\epsilon_l$  and  $\epsilon_h$ ), a method shown to help stabilize the entropy and enhance the model’s exploration capability. The KL divergence loss, controlled by  $\beta$ , is used to prevent the model from deviating excessively from the reference policy distribution.

## B. Experiment Details

### B.1. Training Hyperparameter

We use 8 x A100 GPUs for our experiments in total. GRPO baseline share the same hyperparameters with DoGe rl stage 2. Solver is not learnable, we only provide vllm sampling parameters. Table 5 and 4 present the detailed experimental parameters for base models of 7B and 3B sizes, respectively.

Table 3. 7b-series experiment hyperparameters

Method	Parameter	Value
Thinker	train_steps	150
	train_batch_size	48
	max_prompt_length	768
	max_response_length	4096
	lr	1e-6
	n	4
	temperature	0.9
	clip_ratio_high	0.24
clip_ratio_low	0.2	
Solver	n	4
	temperature	0.9
DoGe Stage 2*	train_steps	150
	train_batch_size	48
	max_prompt_length	768
	max_response_length	4096
	lr	1e-6
	n	8
	temperature	1.0
	clip_ratio_high	0.28
clip_ratio_low	0.2	

Table 4. 3b-series experiment hyperparameters

Method	Parameter	Value
Thinker	train_steps	100
	train_batch_size	64
	max_prompt_length	768
	max_response_length	4096
	lr	1e-6
	n	4
	temperature	0.9
	clip_ratio_high	0.24
clip_ratio_low	0.2	
Solver	n	4
	temperature	0.9
DoGe Stage 2*	train_steps	150
	train_batch_size	64
	max_prompt_length	768
	max_response_length	4096
	lr	1e-6
	n	8
	temperature	1.0
	clip_ratio_high	0.28
clip_ratio_low	0.2	

## B.2. Prompt Template for GRPO Training

In this subsection, we present DoGe’s prompts templates in RL stage 1 and stage 2, where  $\{information\}$  denotes the multimodal reasoning context with the question removed,  $\{Question\}$  denotes the corresponding reasoning question and  $\{deephought\}$  denotes the analysis of Thinker( $\pi_T$ ) to the corresponding question’s contextual information.

### DoGe training stage 1

Thoroughly analyze the provided Contextual Artifacts and Design Information, from which the core question has been deliberately removed.

Your mission is to freely explore, deeply infer, and critically analyze the intrinsic structure, elemental relationships, and underlying constraints of this information.

Generate a detailed, insightful Deep-Thinking Report that will serve as the fundamental guiding knowledge for other models to efficiently solve the original, complete problem.

### Informations  
 $\{information\}$

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within  $\langle think \rangle \langle /think \rangle$  tags. Your Deep-Thinking Report MUST follow  $\langle /think \rangle$  tag.

### Solver Prompt Template

$\{Question\}$  You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within  $\langle think \rangle \langle /think \rangle$  tags. The final answer MUST BE put in  $\backslashboxed{\}$ . If the question is multiple-choice (single- or multi-select), put the final answer inside  $\backslashboxed{\}$ , and format your answer as a Python list of uppercase letters in single quotes, separated by commas (e.g.,  $\backslashboxed{\{['D']\}}$  or  $\backslashboxed{\{['A','B']\}}$ ); otherwise, do not use a list. To solve the problem above, you may refer to the expert analysis of the the given information and the problem scenario.

### Expert Analysis:  
[Analysis Start]  
 $\{deephought\}$   
[Analysis End]

### DoGe training stage 2

$\{Question\}$  You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within  $\langle think \rangle \langle /think \rangle$  tags. The final answer MUST BE put in  $\backslashboxed{\}$ . If the question is multiple-choice (single- or multi-select), put the final answer inside  $\backslashboxed{\}$ , and format your answer as a Python list of uppercase letters in single quotes, separated by commas (e.g.,  $\backslashboxed{\{['D']\}}$  or  $\backslashboxed{\{['A','B']\}}$ ); otherwise, do not use a list.

## B.3. Prompt Template for Problem Synthesis

For each type of disciplinary task, DoGe’s reasoning problem synthesis framework offers two distinct paths: namely, synthesizing problems by sampling the Multimodal Knowledge Pool, and synthesizing variant problems by sampling the Seed Problem PoolS. We respectively denote these two paths as Synthesis 1 and Synthesis 2.  $\{rules\}$  part is optional, and it changes with the generation objective.

### Synthesis 1

You are a knowledgeable and skilled question designer. Design challenging problems based on the given informations and combining your own expertise as well.

The questions you design must follow these rules and format:

$\{rules\}$

Finally, present all questions and answers in a valid JSON block with the following format:

```
““json
[
  {"problem": "imageQuestion text here", "answer": "answer here", "images": ["image path"]},
  {"problem": "Question text here", "answer": "answer here", "images": []}
]
““
```

## Informations  
 $\{information\}$

Let’s think step by step and put your final response in Json Block.

## Synthesis 2

You are a knowledgeable and skilled question designer. You will get some reference seed multimodal problems. Your goal is to design challenging variant problems of given multimodal questions.

The questions you design must follow these rules and format:

{rules}

Finally, present all questions and answers in a valid JSON block with the following format:

```
““json
[
  {"problem": "(image)Question text here", "answer": "answer here", "images": ["image path"]},
  {"problem": "Question text here", "answer": "answer here", "images": []}
]
““
```

## Reference Seed Problems

{examples}

Let's think step by step and put your final response in Json Block.

## B.4. Evaluation Details

Table 5. Evaluation hyperparameters (Backbone Engine: VLLM)

Model Size	Parameter	Value
7B	tensor_parallel_size	2
	gpu_memory_utilization	0.85
	k(mean@k)	4
	max_tokens	4096
	temperature	0.7
3B	tensor_parallel_size	1
	gpu_memory_utilization	0.85
	k(mean@k)	4
	max_tokens	4096
	temperature	0.7

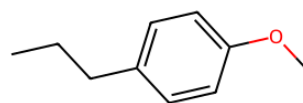
## C. Data Sample

### C.1. Collected Raw Data Examples

As shown in Section 4.1.1, we have collected a large amount of unlabeled data, which serves as the foundation for the learning and evolution of self-evolving VLMs. In this section, we present examples of raw data from the three specific domains(multimodal math, chemistry, earth science).

## Chemistry

### Images



### Raw Text Description

**SMILES:** CCCC1=CC=C(C=C1)OC

**IUPAC:** 1-methoxy-4-propylbenzene

**Formula:** C<sub>10</sub>H<sub>14</sub>O, **MolWt:** 150.2210

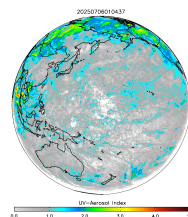
**MolLogP:** 2.6477, **TPSA:** 9.2300

**HBD/HBA:** 0/1, **RotatableBonds:** 3

**RingCount:** 1

## Earth Science

### Images



### Raw Text Description

Image Type: Aerosol; Images returned: 1 of 22 available

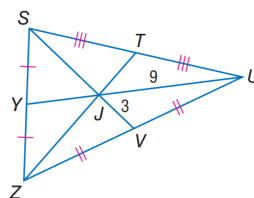
Image 1:Image saved at path1

Date: 2025-xx-xx

Caption: This image was taken by NASA's EPIC camera onboard the NOAA DSCOVR spacecraft

## Math

### Images



### Raw Text Description

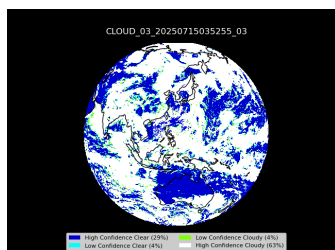
$UJ = 9$ ,  $VJ = 3$ , and  $ZT = 18$ .

## C.2. Synthetic and Masked Problems

We invoke the Gemini2.5-flash-lite to synthesize multi-modal reasoning questions, and conduct masking processing on the original them. specifically, the questions are removed while only the contextual content is retained. We provide some examples below.

### Earth

**Image:**



**Question:**

Consider the cloud classification map. What might be the meteorological significance or interpretation of areas classified as 'Low Confidence Clear' (cyan) or 'Low Confidence Cloudy' (lime green) compared to 'High Confidence Clear' (blue) and 'High Confidence Cloudy' (white)?

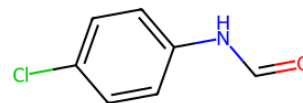
- A. 'Low Confidence' areas typically indicate the presence of very stable atmospheric layers.
- B. 'Low Confidence Cloudy' might suggest thin clouds, haze, or cloud edges, which are harder to classify definitively.
- C. 'Low Confidence Clear' could indicate areas with thin cloud remnants, high humidity haze, or artifacts that don't strongly trigger cloud detection.
- D. These low confidence zones are always associated with immediate precipitation events.
- E. 'High Confidence' classifications indicate areas where the atmospheric conditions are least dynamic.
- F. The low confidence regions often represent transitional zones where atmospheric conditions are changing rapidly.
- G. There is no significant difference in meteorological interpretation between high and low confidence areas.

**Answer:** B, C, F

**Masked Question Contextual Information:**  
cloud classification confidence meaning

### Chemistry

**Image:**



**Question:**

Consider the molecule shown above. What is the sum of its Ring Count and the number of Hydrogen Bond Donors (HBD)?

- A) 1
- B) 2
- C) 3
- D) 4
- E) 5
- F) 6
- G) 7

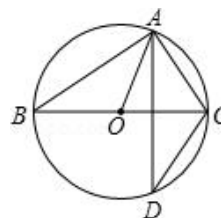
**Answer:** B

**Masked Question Contextual Information:**

Consider the molecule shown above.

### Math

**Image:**



**Question:**

In the given figure, BC is the diameter and O is the center of the circle. If  $\angle ADC = 32^\circ$ , determine the measure of the central angle  $\angle AOB$ .

**Answer:** 116

**Masked Question Contextual Information:**

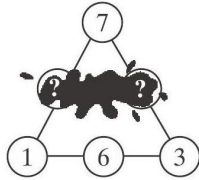
In the given figure, BC is the diameter and O is the center of the circle. If  $\angle ADC = 32^\circ$

## D. Case Study

In this section, we compare the output cases of DoGe, the base model, and the baseline to conduct an in-depth analysis of DoGe's role.

### Example Problem from MathVision

**Image:**



**Question:**

The sums of the all the three numbers on each side of the triangle are equal. Two numbers happened to be stained with ink. How much is the sum of these two numbers?

**Answer:** 2

### Baseline-7B Response.

**Avg Score: 0.0 (mean@4)**

**Response:** *<think>* To solve this, we need to use the information that the sums of the numbers on each side of the triangle are equal. Let's denote the unknown numbers by  $x$  and  $y$ , where  $x$  is the number on the left side (top of the left side), and  $y$  is the number on the right side (top of the right side).

The sums of the numbers on the three sides of the triangle are:

1.  $1 + 6 + 7 = 14$  2.  $7 + 3 + y = 14$

3.  $x + 6 + 3 = 14$

From the second equation, we can solve for  $y$ :

$$7 + 3 + y = 14$$

$$10 + y = 14$$

$$y = 4$$

From the third equation, we can solve for  $x$ :

$$x + 6 + 3 = 14$$

$$x + 9 = 14$$

$$x = 5$$

Now, we need to find the sum of  $x$  and  $y$ :

$$x + y = 5 + 4 = 9$$

Thus, the sum of the two numbers  $x$  and  $y$  is:

*</think>* 9

### DoGe-7B Response.

**Avg Score: 0.5 (mean@4)**

**Response:** *<think>* To solve this problem, we start by using the given condition that the sums of the numbers on each side of the triangle are equal. Let's denote the numbers at each of the three points as follows: the top point as 7, the left point as 1, the right point as 3, and the unknown numbers at the bottom as  $x$  and  $y$ .

1. The sum of the numbers on the left side is:  
 $1 + 6 + x$ .

3. The sum of the numbers on the top side is:  
 $7 + y + x$ .

Since the sums are equal, we can set these expressions equal to each other:

$$1 + 6 + x = 3 + 6 + y = 7 + y + x$$

From the first two sides, we can set them equal and solve for  $x$  and  $y$ :

$$1 + 6 + x = 3 + 6 + y \implies 7 + x = 9 + y$$

$$\implies x - y = 2$$

From the second and third sides, we can set them equal and solve for  $x$  and  $y$ :

$$3 + 6 + y = 7 + y + x \implies 9 + y = 7 + y + x$$

$$\implies 9 = 7 + x \implies x = 2$$

Now, substituting  $x = 2$  back into the first equation:

$$2 - y = 2 \implies y = 0$$

So, the numbers are  $x = 2$  and  $y = 0$ . The sum of these two numbers is:

$$x + y = 2 + 0 = 2$$

*</think>* The sum of the two stained numbers is 2.

**Avg Score: 0.0 (mean@4)**

**Response:** To solve the problem, we need to determine the values of the two numbers that are stained with ink in the triangle. The sums of the numbers on each side of the triangle must be equal.

Let's denote the numbers as follows: - Top of the triangle: 7 - Left side: 1, 6, and  $x$  (unknown number) - Right side: 3,  $y$  (unknown number), and  $z$  (unknown number)

The sums of the numbers on each side of the triangle are equal. Therefore, we can set up the following equations:

$$x - z = -3$$

$$z = x + 3$$

Now, we need to find the value of  $y$  and  $z$  in terms of  $x$ :

$$y = x + 3$$

$$z = x + 3$$

Since the sums of the numbers on each side must be equal, we can check the sum of the left side:

$$7 + 6 + x = 13 + x$$

And the sum of the right side:

$$7 + 3 + z = 10 + z$$

Since  $z = x + 3$ :

$$10 + (x + 3) = 13 + x$$

This equation is always true, so the values are consistent. Now, we need to find the sum of the two numbers that are stained with ink, which are  $y$  and  $z$ :

$$y + z = (x + 3) + (x + 3) = 2x + 6$$

To find  $x$ , we can use the sum of the numbers on the left side:

$$7 + 6 + x = 13 + x$$

Since the sums are equal, we can choose any consistent value for  $x$ . Let's assume  $x = 0$ :

$$y = 0 + 3 = 3$$

$$z = 0 + 3 = 3$$

So, the sum of  $y$  and  $z$  is:

$$y + z = 3 + 3 = 6$$

Therefore, the sum of the two numbers that are stained with ink is:

6

## E. Prompt Template for Evaluation

### LLM as Judge Prompt

### Evaluation Criteria (Score 1-10) Evaluate the report based on the following criteria. \*\*Do not penalize for Markdown rendering errors or structural formatting issues.\*\* Focus solely on the content and reasoning quality.

- Perception & Coverage**: Does the report accurately identify all key entities in the text and images (if present)? (e.g., interpreting SMILES strings correctly, recognizing visual features).
- Depth of Inference**: Does the report go beyond surface-level descriptions? Does it infer underlying properties, potential constraints, or causal relationships? (e.g., calculating molecular weight, inferring solubility based on structure, distinct feature comparison).
- Logic & Coherence**: Is the internal monologue and analysis logically sound? Are there contradictions?
- Utility for Downstream Tasks**: Since the core question is missing, does this report provide a "comprehensive map" that would allow another model to answer *any* likely question about this data?

### ### Scoring Scale

- 1-2 (Poor)**: Hallucinations, factually incorrect, or misses the vast majority of information.
- 3-4 (Weak)**: Only describes the surface level (e.g., "This is a molecule") without analysis.
- 5-6 (Average)**: Accurate description but lacks deep inference or cross-modal synthesis.
- 7-8 (Good)**: Strong analysis, correct inferences, connects different pieces of information well.
- 9-10 (Excellent)**: Exceptional insight, uncovers hidden relationships, highly predictive, and serves as a perfect knowledge base.

### ### Output Format

You must strictly output a JSON object with the following structure:

```

{
  "reasoning": "A concise explanation of the pros and cons of the report based on the criteria.",
  "score": <int>
}

```