

## 424 A Technical Appendices and Supplementary Material

### 425 A.1 Training and Evaluation Details

Table 4: Details of evaluation benchmarks.

Benchmark	Description	#samples
Mantis-eval	Multi-image General Understanding QA	217
BLINK	Multi-image General Understanding QA	1901
MMIU	Multi-image General Understanding QA	11698
MathVista	Single-image Math Reasoning QA	1000 (testmini)
MathVerse	Single-image Math Reasoning QA	3940
MathVision	Single-image Math Reasoning QA	3040
Remi	Multi-image General Reasoning	2600
MV-Math	Multi-image Math Reasoning	2009

426 We train our model on 8×H100 GPUs using the GRPO-based framework. The training batch size is  
427 128, with a micro-batch size of 4 per GPU. The maximum prompt and response lengths are 8192 and  
428 2048 tokens, respectively. We adopt Qwen2.5-VL-7B-Instruct<sup>2</sup> as the backbone, enabling gradient  
429 checkpointing and FSDP without offloading. KL loss is applied with a coefficient of 0.01, while  
430 entropy regularization is disabled. The model is optimized for 2 epochs with a learning rate of  
431  $1 \times 10^{-6}$ .

432 For evaluation, we use greedy decoding with temperature set to 0, top-p to 1, top-k to -1, and  
433 a maximum generation length of 2048. Evaluation benchmarks include MathVista, MathVerse,  
434 MathVision, and BLINK (configured via VLMEvalKit), while Mantis-Eval, MMIU, and MV-MATH  
435 (evaluated with official code via vLLM). All evaluations follow consistent decoding settings. Details  
436 are shown as Table 4 and prompt A.3

### 437 A.2 Case study

438 As shown in Fig 6, we conduct a detailed comparison of the response quality across different types  
439 of models, including the proprietary model o4-mini-high, open-source base models, and previously  
440 trained interleaved models. We observe that the base models exhibit various types of errors, which  
441 can be broadly categorized into several patterns: missing relevant images, misinterpreting images,  
442 producing incorrect answers despite correct reasoning, and reaching incorrect conclusions despite  
443 accurate observations of details. For instance, in the beetle question shown above, Qwen frequently  
444 makes errors where the reasoning is correct but the final answer is wrong, or where some relevant  
445 visual information is overlooked. These are the types of failures we aim to address.

446 Besides, in case fig 8, we observe a consistent pattern in model behavior across the permutations of  
447 images: the model exhibits a stronger alignment between the reference image and the image presented  
448 first in the sequence of options. This suggests current VLMs have a positional bias in interleaved  
449 vision tasks.

<sup>2</sup><https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct>

Question: How many beetles are climbing on the balls in the image?

A. Four B. Five C. None of the choices provided D. Two E. One



Qwen 2.5 VL: <thinking> The image shows two beetles actively engaged in rolling a ball of dung. There are no other beetles visible on the balls or in the immediate vicinity. Therefore, the number of beetles climbing on the balls is two.</thinking> <answer> D. Two </answer>

LLAVA-Interleave: <think> The left image shows a beetle on a ball. The right image shows two beetles. Therefore, the number is three. </think> <answer> D.Two</answer>

o4-mini-high: <thinking>In the first image there are two beetles on the ball. In the second image there is one beetle on the ball. That makes three beetles in total, and three is not one of the listed options.</thinking> <answer>None of the choices provided</answer>

Ours: <thinking>...<thinking> <answer>None of the choices provided</answer>

Figure 6: A showcase for motivation

Here is a statement describing these 2 images: The train in one of the images has just come around a bend. Is it true or false?

(A) False  
(B) True

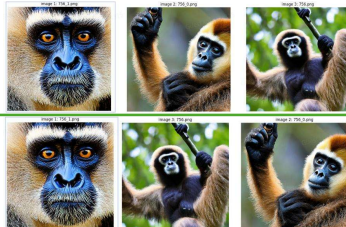


Which star is green?

A. <image> B. <image> C. <image>



If image 3 is the reference image, which image of the other two is more similar to the reference image? Image 2 Image 3



Could you predict the series of actions in this video and enumerate them in order. <image> <image> <image> <image> <image> <image> . options:

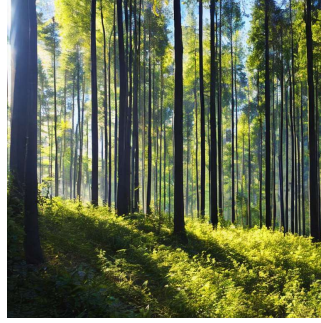
A. 'reach bowl, hold fork, transfer soup, carry cup'  
B. 'hold cup, reach cup, carry cup'  
C. 'flip cup, hold plate, pour juice, carry spoon'  
D. 'None of the choices provided'



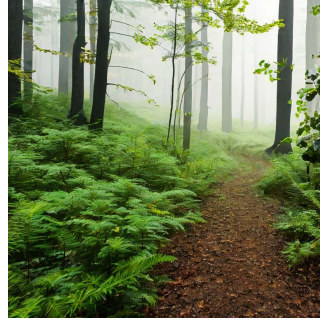
Figure 7: Training Data preprocessing pipeline explanation and several cases



<image\_1>



<image\_2>



<image\_3>

### Before Permuation

Which image is more similar to the reference image <image\_1>? Options: A. <image\_2> B. <image\_3>

**Qwen-2.5-VL-7B:** To determine which image is more similar to the reference image, let's analyze the key elements of both images:

- Reference Image: 1) Features tall trees with a dense forest setting. 2) The lighting appears to be natural, possibly early morning or late afternoon, with sunlight filtering through the trees.
- Option A: 1) Also shows a forest scene with tall trees. 2) There is less mist or fog compared to the reference image.
- Option B: 1) Displays a forest path surrounded by lush greenery. 2) There is no mist or fog present, and the overall atmosphere is more open and less dense than the reference image.

Comparing these elements, Option A is closer to the reference image because it maintains the tall trees and dense forest setting. ✓

### After Permuation

Which image is more similar to the reference image <image\_1>? Options: A. <image\_3> B. <image\_2>

**Qwen-2.5-VL-7B:** To determine which image is more similar to the reference image, let's analyze the key elements of both images:

- Reference Image: 1) Features tall trees with a dense forest setting. 2) The lighting suggests early morning or late afternoon, with sunlight filtering through the trees. 3) The ground is covered with green foliage, and there is a misty atmosphere.
- Option A: 1) Also features a forest with tall trees. 2) There is a path visible, which is not present in the reference image. 3) The lighting appears to be diffused, possibly due to the mist, but it is less pronounced than in the reference image. 4) The ground is covered with ferns and other greenery, similar to the reference image.
- Option B: 1) Features tall trees with a dense forest setting. 2) The lighting is bright and direct, suggesting midday sun. 3) The ground is covered with green foliage, but the overall scene lacks the misty atmosphere seen in the reference image.

Given this analysis, option B lacks the misty atmosphere and has brighter lighting, making it less similar to the reference image. The image that is most similar to the Reference Image is Option A. ✗

Figure 8: Case study demonstrating positional bias. The model correctly identifies the similar image when presented first (Before Permuation) but fails when the order is swapped (After Permuation).

## 450 A.3 prompt

### 451 A.3.1 Prompt for data preprocessing

452 During the data preprocessing stage, we employed GPT-4o in conjunction with the prompts provided  
453 below to reformat questions and perform semantic variation checks.

#### Format Rephrase

Your task is to convert a given QA conversation into a multiple-choice question format, and determine whether the question follows the given question type.

#### Question Type Definition

1) A question is considered a **Reference-Image Comparison** if it satisfies all of the following conditions:

- The question presents three or more images (e.g., "<image> <image> <image>").
- One image is clearly identified as the **reference image** (e.g., "image 1 is the reference").
- The question asks which of the remaining images is most similar to or most different from the reference image.
- The answer options correspond only to the non-reference images.

2) ...

#### Task Instructions

1) **Convert the original QA pair into a multiple-choice question:**

- Rephrase the assistant's response into an answer option (e.g., "A", "B", "C").
- Use placeholder tokens (<image>) in both the question and the options.
- Include only the images being compared (exclude the reference image from the options).
- Format the question strictly according to the example below.

2) **Determine the type of the question.**

#### Output Format

Return a JSON object with the following structure:

```
{
  "question": "<multiple-choice question in specified format>",
  "answer": "<correct option letter>",
  "question_type": "<ReferenceComparison or Other>"
}
```

#### Example

**Input:** "Question: Answer the following question: Here are three images: <image\_1> <image\_2> <image\_3>. If image 1 is the reference image, which image of the other two is more similar to the reference image? Answer: The image that is more similar to the reference image is image 2."

**Output:**

```
{
  "question": "Answer the following question: Which image
    is more similar to the reference image <image_1>?
    Options: A. <image_2> B. <image_3>",
  "answer": "A",
  "question_type": "ReferenceComparison"
}
```

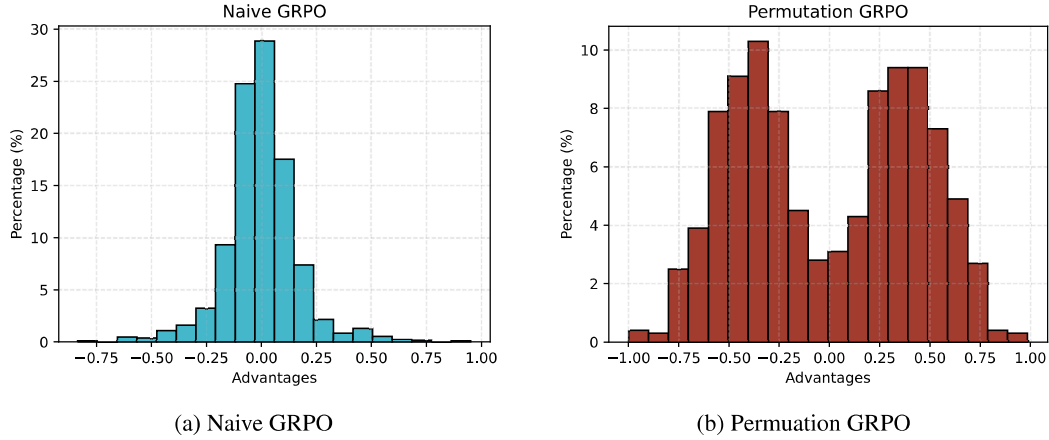


Figure 9: Difference on advantage between Naive GRPO and Permutation GRPO during training stage.

### Semantic Variation Check

You will be given a question involving one or more images, which are represented using image tokens. The token `<image>` represents an actual image. The tokens `<image_1>`, `<image_2>`, etc., refer to specific images by their positions (e.g., `<image_1>` refers to the first image, `<image_2>` to the second image).

**Your task is to determine the following:**

1. If only the order of the images (e.g., `<image_1>`, `<image_2>`, etc.) is changed, would the answer to the question need to change?
2. Is the question structured such that a single main image appears in the question body, and other images are referenced in the choices?

**Respond in the following JSON format:**

```
{
  "should_change": true or false,
  "is_multichoice_images": true or false
}
```

*Note: This applies to both multiple-choice and fill-in-the-blank questions involving image references.*

### A.3.2 Prompt for training and evaluation

#### Reasoning Format

##### Instruction:

You first think about the reasoning process as an internal monologue and then provide the final answer.

The reasoning process must be enclosed within `<think>` `</think>` tags.

The final answer must be put in `\boxed{}`.

### A.4 Advantage Differences

As Fig 9 shows, the distinction becomes clearer when examining the advantage distributions during training. For multi-image inputs, Naive GRPO yields a distribution sharply peaked around zero, indicating that a large portion of training examples contribute negligible or ineffective gradient signals.

462 In contrast, the permutation-based GRPO introduces greater input diversity by altering image order,  
463 which encourages the model to genuinely capture positional biases. This diversification leads to more  
464 informative advantage signals, facilitating more effective gradient updates during optimization.