
Supplementary Material For ‘Actial: Activate Spatial Reasoning Ability of Multimodal Large Language Models’

Anonymous Author(s)

Affiliation

Address

email

1 We analyze the training process in Section. A. We further present dataset examples and the utilized
2 prompts in Section. B and Section. C, respectively.

3 A Training Analysis

4 Figure 1(a) presents the evolution of the model’s response length throughout the GRPO phase. Our
5 observations align with those reported in [2], where Actial displays an initial reduction followed
6 by a subsequent adjustment in response length. When contrasted with direct fine-tuning of the
7 baseline model, Actial demonstrates an extended reasoning length. The green line (without hybrid
8 cold-start initialization) reflects no significant increase in response length, attributable mainly to the
9 Supervised Fine-Tuning (SFT) stage. This stagnation can be explained by the composition of our
10 Viewpoint-100K dataset, which comprises exclusively multiple-choice questions devoid of reasoning
11 templates. Consequently, the model struggles to accurately achieve format rewards during the GRPO
12 process, and we introduce the hybrid cold-start initialization to improve such issue.

13 Figure 1(b) depicts the evolution in KL Divergence across different variants. The green line, which
14 represents the direct application of Viewpoint-100K for GRPO, exhibits a growing offset relative
15 to the initial strategy, suggesting that the baseline model’s original spatial reasoning capabilities
16 are inadequate for handling viewpoint-specific tasks. Conversely, utilizing the SAT dataset leads
17 to substantially lower KL divergence, underscoring the unique spatial reasoning demands posed by
18 our Viewpoint-100K dataset. Following knowledge injection, the KL divergence becomes notably
19 smoother, indicating the efficacy of integrating foundational viewpoint knowledge.

20 Figure 1(c) illustrates the pairwise loss observed during the Supervised Fine-Tuning (SFT) training
21 phase. Notably, when using our Viewpoint-100K dataset, there is a sudden and significant improve-
22 ment in performance (a trend also reflected in the validation curve). This rapid decrease excludes the
23 possibility of the model simply memorizing the answers since it appears within one single epoch.
24 Additionally, when using a hybrid cold-start initialization, which requires the model to learn reasoning
25 templates, the loss curve becomes more smoother. However, the sudden insight remains clearly
26 evident, reflecting its relevance to our viewpoint-based questions. We believe that this phenomenon
27 mainly comes from two reasons. First, our dataset consists of relatively simple multiple-choice
28 questions with only three options (one of which is a distractor), making it easy for the model to select
29 the correct answer even without proper reasoning. However, such correctness achieved through flawed
30 reasoning is insufficient for the model to truly understand and solve viewpoint-related problems,
31 leading to a oscillation phase of loss fluctuation. Second, as mentioned in the main text, existing
32 MLLMs tend to rely on incorrect 2D cues when solving 3D tasks. In contrast, viewpoint problems
33 require the model to learn how to properly utilize 3D spatial cues, resulting in a period where the
34 loss remains stagnant. Nevertheless, MLLMs do possess latent 3D perception capabilities. Once
35 the model learns to shift its perspective appropriately (seems to be like the activation), it can rapidly
36 generalize this understanding to similar tasks, leading to a sudden drop in loss.

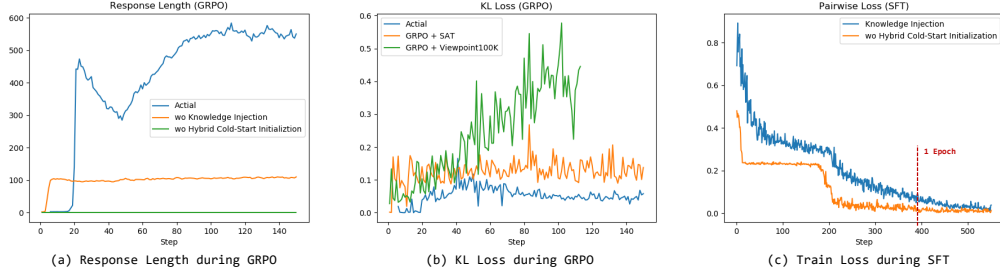


Figure 1: Metrics changes during the training process.

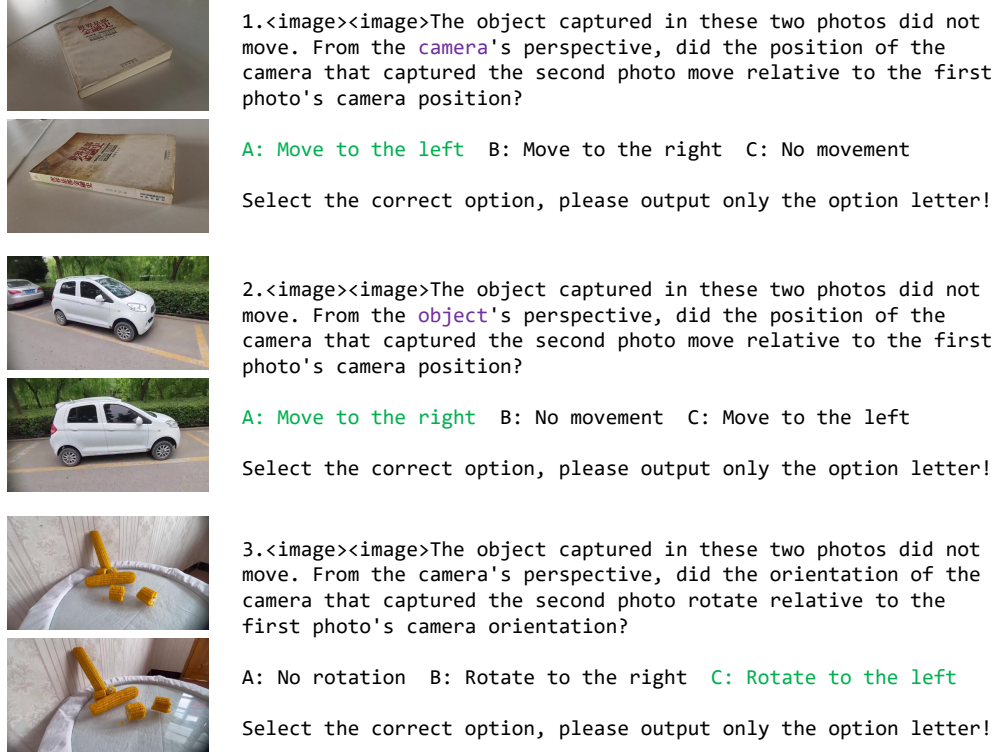


Figure 2: The QA examples of Viewpoint-100K. We provide three types of questions.

37 B Dataset Examples

38 We show the examples of Viewpoint-100K in Figure. 2. We provide three types of questions,
 39 including the horizontal translation and rotation from the camera's perspective and the horizontal
 40 translation from the object's perspective. We also provide the accurate rotation angles in our dataset,
 41 calculated from the camera parameters.



Gemini-2.5-Pro

```
<think>
Okay, let's analyze the camera movement based on the two images provided.

**1. Identify the Images:**
*Image 1: Shows an open book placed on a wooden shelf against a wood-paneled wall. The left page has minimal content, while the right page is partially obscured but shows some red lantern illustrations near the top. A gold-colored object (phone/power bank) props the book open. A sign ("温馨提示...") is visible on the wall to the right.*
*Image 2: Shows the same book, shelf, wall, and gold object. The right page is now fully visible, revealing a detailed illustration. The sign on the wall is no longer in view, but a power outlet is now visible further to the right on the wall.*

**2. Compare the Images - Visual Analysis:**
**From Camera's Perspective**
*Framing:* In Image 1, the book occupies the center-left portion of the frame. The sign on the wall is clearly visible on the right. In Image 2, the book has shifted significantly to the left side of the frame. The sign is gone, and the power outlet, which must be located further right than the sign, is now included in the frame on the far right.*
*Perspective/Angle:* In Image 1, the camera views the book slightly from the left. The left page is more open to the camera, while the right page is viewed at a more oblique angle. We see the front edge and the right side (spine edge) of the book block. In Image 2, the camera views the book more from the right. The right page is now much more open and directly facing the camera, while the left page is viewed more obliquely. The front edge and right side of the book block are seen from a different viewpoint.*
*Scale/Distance:* The book appears slightly larger in Image 2 compared to Image 1, suggesting a potential slight forward movement of the camera.*
*Background:* The change in visible background elements (sign disappears, outlet appears) confirms a significant horizontal shift in the camera's position relative to the wall.*

**From Object's (Book's) Perspective:**
*Camera Position:* In Image 1, the camera is positioned somewhat to the left of the book's spine and slightly above it. In Image 2, the camera is positioned significantly to the right of the book's spine and slightly above it.*
*Visible Surfaces:* The change in which page (left vs. right) is more directly facing the camera indicates a change in the camera's viewing angle relative to the book's orientation.*

**3. Analyze Specific Movements:**
**Left/Right Translation:** The most apparent change is the shift of the book to the left within the frame and the appearance of background elements (power outlet) further to the right. This indicates the camera physically moved **right** relative to the scene between shots.*
**Up/Down Translation:** The vertical perspective seems very similar. There's no strong indication of significant up or down movement. We can consider this negligible.*
**Front/Back Translation:** The book appears slightly larger in Image 2, suggesting the camera moved slightly **forward** (closer to the book).*
**Rotation (Yaw - around vertical axis):** In Image 1, the camera aims slightly towards the right relative to the book's spine (capturing the left page more frontally). In Image 2, the camera aims significantly towards the left relative to the book's spine (capturing the right page more frontally). To achieve this while translating right, the camera must have rotated **left** (counter-clockwise when viewed from above) around its vertical axis.*
**Rotation (Pitch - up/down tilt):** Minimal change observed. Any change is likely negligible.*
**Rotation (Roll - side tilt):** The horizontal lines of the shelf and wall appear consistently parallel to the image bottom edge. No significant roll detected.*

**4. Synthesize the Trajectory:**
From taking Image 1 to Image 2, the camera primarily moved **right** and slightly **forward**. Simultaneously, it rotated **left** around its vertical axis (yaw). Up/down translation, pitch, and roll were minimal.

**5. Answer Question 1 (Horizontal Translation - Camera Perspective):**
From the camera's perspective, comparing its position in the second shot relative to the first, the camera moved horizontally to the **right**.

**6. Answer Question 2 (Horizontal Translation - Object Perspective):**
From the object's (book's) perspective, imagine standing on the book facing forward (away from the wall). The camera, which was initially somewhat to your left (in Image 1), moved to be significantly to your right (in Image 2). Therefore, from the object's perspective, the camera's horizontal translation relative to the object was to the **left**.

**7. Answer Question 3 (Rotation around Vertical Axis - Camera Perspective):**
The camera rotated around its own vertical axis relative to its orientation in the first shot.
Direction: To change the view from favoring the left page (Image 1) to favoring the right page (Image 2) while physically moving right, the camera had to rotate **left** (counter-clockwise).
Estimate the angle: The change in perspective on the book's pages is significant. In Image 1, the camera is perhaps looking 20-30 degrees to the right of the book's spine normal. In Image 2, it's looking perhaps 20-30 degrees to the left of the spine normal. The total change in angle relative to the book is substantial. Estimating this change based on how much more "open" the right page becomes suggests a rotation of approximately **54 degrees** to the left.
</think>
<answer>A</answer>
```

Figure 3: An example of our generated pseudo CoT.

42 C Prompts

43 Our system prompt is copied from VLAA [1], shown as follow:

System Prompt:

You are Actial, a helpful assistant with excellent reasoning ability.

A user asks you a question, and you should try to solve it.

You should first think about the reasoning process in the mind and then provides the user with the answer.

The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>"

44

45 We show an example of our generated pseudo CoT in Figure. 3. The human-assisted CoT template is
46 as follow:

Human-Assisted CoT Template:

Based on the content presented in the two pictures, analyze the movement trajectory of the camera from shooting the first picture to shooting the second one, Describe the movement trajectory specifically, including front, back, left, right, up and down, as well as rotation . and then answer the following two questions based on the camera trajectory.

1. From the camera's perspective, determine the horizontal translation of the second camera relative to the first camera.
2. From the object's perspective, determine the horizontal translation of the second camera relative to the first camera.
3. From the camera's perspective, determine the rotation of the second camera around its own verticalaxis relative to the first camera's orientation:
 - Estimate the rotationangle in degrees.

Thinking:

I need to get the relevant content to answer the user's question.

Here's a thinking process to analyze the camera movement between the two images:

Understand the Goal: The user wants to know how the camera moved between taking the first picture (1.jpg) and the second picture (2.jpg). The object did not move in any way from the first picture to the second. The description needs to cover translation (front/back, left/right, up/down) and rotation. Then, I need to answer two specific questions: horizontal translation (left/right/none) and rotation angle around the camera's vertical axis.

1. Identify the Images:

Image 1: 1.jpg (Let's call this Pic 1)

Image 2: 2.jpg (Let's call this Pic 2)

The object in both pictures is a red, quilted handbag with a gold chain strap, placed on a blue-and-white striped surface with yellow stars and pink crescents.

2. Compare the Images - Visual Analysis:

From camera's perspective:

Frames:

In Pic 1, we see more of the surface above the bag. From a camera centered perspective, we can see the left plane of the bag, but we cannot see the right plane of the bag.

In Pic 2, from a camera-centered perspective, the front of the bag is visible, with the golden

47

chain strap and logo pointing toward the bottom of the image. The left side is not visible.

Perspective/Angle:

In Pic 1, from the camera's perspective, shows more of the front and the left side of the bag. The chain strap is clearly visible and faces towards the bottom left of the photo.

In Pic 2, from the camera's perspective, shows more of the front side of the bag. The clasp is still visible but from a slightly different angle, it faces towards the bottom of the photo. The top surface of the bag seems a bit more visible in Pic 2, suggesting a slightly higher viewpoint or a slight downward tilt.

From object's perspective:

Frames:

In Pic 1, define the side of the bag featuring the gold chain strap and logo as its front. From the bag's perspective, the camera is positioned slightly to the left and above the bag, with a viewpoint angled downward and toward the left plane of the bag.

In Pic 2, from the bag's perspective, the camera moves to the right. It is now positioned slightly in front of and above the bag, with a slight downward angle.

Perspective/Angle:

In Pic 1, from the bag's perspective, both the front and left planes are visible to the camera.

In Pic 2, from the bag's perspective, the left plane is no longer visible, indicating that the camera has rotated to face more directly in front of the bag.

Scale/Distance: The bag appears roughly the same size in both images, suggesting minimal movement towards or away from the object (front/back translation).

Background: The wall/surface behind the striped fabric looks similar in both, although the angle relative to the bag changes.

3. Analyze Specific Movements:

Left/Right Translation: If the camera undergoes only horizontal translation, we can compare the position of the bag relative to the image frame. However, in this case, the camera trajectory involves rotation as well. In Pic 1, both the front and left planes of the bag are visible. In Pic 2, the left plane is no longer visible, while the front plane remains in view. The change in visible surfaces indicates that the camera must have moved to the left relative to the bag.

Up/Down Translation: In Pic 1, there's more space above the bag. In Pic 2, there's less space above the bag, and the camera orientation seems slightly higher relative to the bag's top surface. This suggests the camera moved slightly up.

Front/Back Translation: The scale of the bag seems very similar. There's no strong indication of significant movement towards or away from the bag. Let's assume minimal or no front/back translation.

Rotation (Yaw - around vertical axis): In Pic 1, the front plane of the bag faces toward the bottom-left of the photo. In Pic 2, the front plane faces toward the bottom of the photo. Given the leftward horizontal translation of the camera, to achieve this change in perspective while keeping the bag relatively stationary, the camera must have rotated rightward (clockwise when viewed from above) around its vertical axis.

Rotation (Pitch - up/down tilt): The top surface seems slightly more visible in Pic 2. This could be due to the upward translation or a slight downward tilt (pitch down). Given the upward translation seems likely, a slight downward tilt might also be present, but the upward movement is more apparent. Let's focus on the dominant movements.

Rotation (Roll - side tilt): The horizontal lines of the striped fabric and the wall seem roughly parallel to the image edges in both pictures. There's no significant roll.

4.Synthesize the Trajectory:

From Pic 1 to Pic 2, the camera moved primarily left and slightly up. It also rotated rightward around its vertical axis (yaw). There was likely minimal front/back movement and minimal roll or pitch changes.

5.Answer Question 1 (Horizontal Translation in camera's perspective):

The camera moved horizontally relative to the bag.

Did it move left or right? It moved left.

The question asks for the horizontal translation of the second camera relative to the first camera. The answer is left.

6.Answer Question 2 (Horizontal Translation in object's perspective):

Consider the object's perspective, the camera now moved to the right. Did it move left or right? It moved right.

The question asks for the horizontal translation of camera in object's perspective.

The key point of the problem is to transform the reference frame onto the object. The answer is left.

7.Answer Question 3 (Rotation around Vertical Axis in camera's perspective):

The camera rotated around its vertical axis relative to its orientation in the second shot.

Did it rotate left or right? It rotated rightward.

Estimate the angle. Looking at the changes in visible parts of objects in the two photos, a significant rotation occurred. Comparing the visible faces, the first photo includes the front and left part of the bag in the camera's perspective. In Pic 2, we see the front almost squarely, maybe slightly revealing the right edge. It looks like roughly a 80 to 90-degree change in viewing angle relative to the bag.

Did it rotate to the left or right? Since the original left part of the bag in the first photo can not be seen in the second photo, it represents the camera rotates to the right, resulting in more visible part of the left side of the front of the bag.

Estimate: Let's go with approximately 87 degrees. (This is an estimation based on the visual change). The answer is the camera rotates 87 degrees to the right.

7.Format the Output: Combine the trajectory description and the answers to the specific questions clearly. Ensure the answers use the specified format.

49

50 The prompt used to generate the pseudo CoTs is as follow:

Pseudo CoT Generation:

<image><image>

I'll give you an example, please follow the example to analyze the camera movement between the two images.

[*Human-Assisted CoT Template*]

Please analyze the given two images using the same approach as I provided above. Including the Image Identify, Visual Analysis, Analyze Specific Movements, Synthesize the Trajectory, Answer Question.

Do not use the relative position changes of objects in the image, consider the changes in three-dimensional space.

The correct answer is: move to the *gt[0]* in camera's perspective; move to the *gt[1]* in object's perspective; rotate to the *gt[2]* with about *gt[3]* degrees. Please provide a sufficiently detailed analysis..

51

References

- 53 [1] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft
54 or rl? an early investigation into training rl-like reasoning large vision-language models, 2025.
- 55 [2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong
56 Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement
57 learning. *arXiv preprint arXiv:2501.12948*, 2025.