

# ViSS-R1: Self-Supervised Reinforcement Video Reasoning

## Supplementary Material

### A. Analysis of Visual Underutilization

Our motivation for visual-centric video reasoning via self-supervised reinforcement learning stems from the observation that substantial video information is **underutilized** in previous video R1 methods. To demonstrate this issue, we design an interesting experiment as shown in Fig. 7. During inference, we randomly mask frames at specific ratios in the test videos and measure the performance drop percentage of the Video-R1 model compared to results without masking. A higher masking ratio indicates that the video sequences are more likely to be perturbed, missing information and less sequential, posing greater challenges for accurate reasoning.

As shown in Fig. 7, the previous Video-R1 [14] model is generally insensitive to frame masking, exhibiting only minor performance drops, especially on reasoning benchmarks such as VSI-Bench and MMVU. Notably, on MMVU with 40% of frames masked, Video-R1 even shows improved reasoning results (i.e., negative drop percentage). These findings indicate that visual information is not fully utilized in the Video-R1 model, as masking out visual frames has little impact on overall reasoning performance (on a visual task). In contrast, our ViSS-R1 model is more sensitive to visual frame masking, demonstrating that more visual knowledge is actually leveraged during reasoning.

### B. Pretext Question Prompts

We provide detailed and specific prompts for pretext questions used in Pretext-GRPO. The pretext question for the image rotation task is shown in Fig. 8, for the image flip task in Fig. 9, and for the image puzzle task in Fig. 10. The pretext question for the video rotation task is shown in Fig. 11, for the video arrow task in Fig. 12, and for the video shuffle task in Fig. 13.

### C. More Qualitative Results

We present additional qualitative results from several challenging benchmarks, including VSI-Bench (Fig. 14), VideoMMU (Fig. 15), and VideoMME (Fig. 16). For a fair comparison between Video-R1 and our ViSS-R1, we employ the same inference prompt for both models. Detailed analyses of specific videos and questions are provided in the respective figure captions.

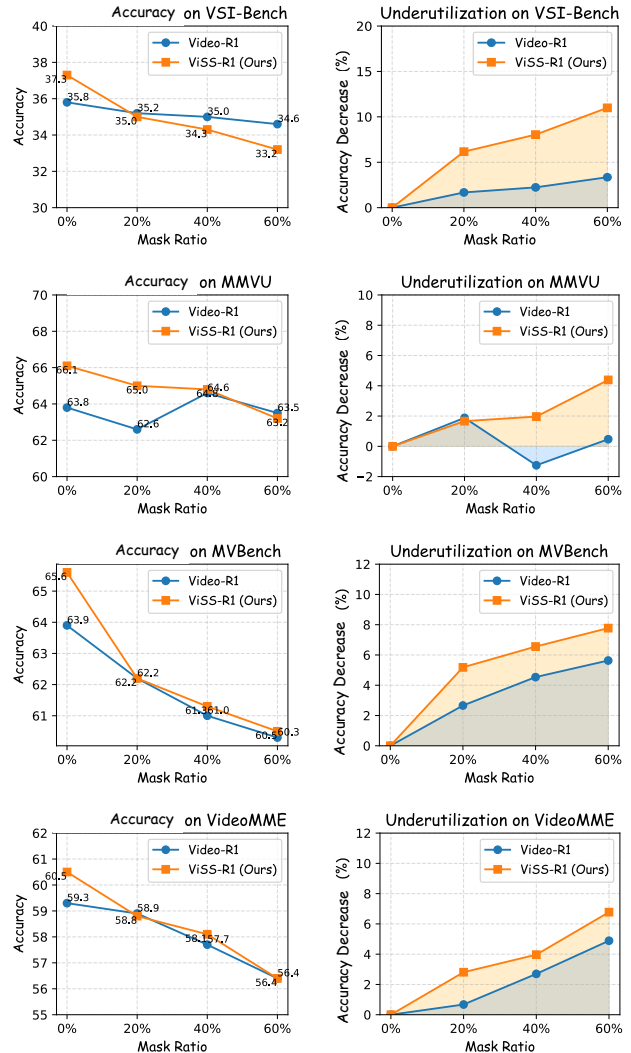


Figure 7. **Illustration of underutilization problem.** (left) accuracy vs. frame masking ratio. (right) relative decrease in accuracy vs. frame masking ratio. The performance of Video-R1 does not decrease much when visual frames are masked out, indicating under-utilization of visual information during reasoning. In contrast, the accuracy of ViSS-R1 decreases more when frames are masked out, showing its reliance of visual frames during reasoning.

### Pretext Question of Image Rotation

Please review the transformed image and determine which rotation angle was most likely applied to the original image. The possible options are:

- A.  $0^\circ$
- B.  $90^\circ$
- C.  $180^\circ$
- D.  $270^\circ$

Note: The rotation is applied in the counterclockwise (anticlockwise) direction and  $0^\circ$  corresponds to no rotation. Please focus on the spatial cues of the image and think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Figure 8. Image rotation prompt.

### Pretext Question of Image Puzzle

Please carefully review the transformed image. The image has been evenly divided into four equal-sized patches, labeled as follows:

- Patch 1: Top-Left
- Patch 2: Top-Right
- Patch 3: Bottom-Left
- Patch 4: Bottom-Right

Two of these patches have been randomly selected and swapped with each other, while the other two patches remain in their original positions. Your task is to determine which pair of patches were swapped.

Here are your options:

- A. Patch 1 and Patch 2 were swapped
- B. Patch 1 and Patch 3 were swapped
- C. Patch 1 and Patch 4 were swapped
- D. Patch 2 and Patch 3 were swapped
- E. Patch 2 and Patch 4 were swapped
- F. Patch 3 and Patch 4 were swapped

Please focus on the spatial details and cues within the image contents and think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags. Please provide only the single option letter (e.g., A, B, C, D, E etc.) within the <answer> </answer> tags.

Figure 10. Image puzzle prompt.

### Pretext Question of Image Flip

Please review the input image and determine whether the image has been flipped.

The possible options are:

- A. No, the image remains in its original position.
- B. Yes, the image has been flipped '\horizontally\'.
- C. Yes, the image has been flipped '\vertically\'.

Please focus on the image contents and think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Figure 9. Image flip prompt.

### Pretext Question of Video Rotation

Please review the transformed video clip and determine which rotation angle was most likely applied to the original video.

The possible options are:

- A.  $0^\circ$
- B.  $90^\circ$
- C.  $180^\circ$
- D.  $270^\circ$

Note: The rotation is applied in the counterclockwise (anticlockwise) direction and  $0^\circ$  corresponds to no rotation. Please focus on the spatial cues of video contents and think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Figure 11. Video rotation prompt.

### Pretext Question of Video Arrow

Please review the input video clip and determine whether the sequence of video frames has been reversed (i.e., played in reverse order).

The possible options are:

- A. No, the video frames remain in their original order.
- B. Yes, the video frames have been reversed.

Please focus on the temporal order of video contents and think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Figure 12. Video arrow prompt.

### Pretext Question of Video Shuffle

Please review the input video clip. The video originally consists of 16 frames, divided into 4 consecutive groups as follows:

Group 1: Frames [1, 2, 3, 4]

Group 2: Frames [5, 6, 7, 8]

Group 3: Frames [9, 10, 11, 12]

Group 4: Frames [13, 14, 15, 16]

Now two of these groups have been randomly selected and swapped with each other, while the remaining groups stayed in their original positions. Please identify which two groups are swapped.

- A. Group 1 and Group 2 were swapped
- B. Group 1 and Group 3 were swapped
- C. Group 1 and Group 4 were swapped
- D. Group 2 and Group 3 were swapped
- E. Group 2 and Group 4 were swapped
- F. Group 3 and Group 4 were swapped

Please focus on the temporal continuity of video contents and think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Figure 13. Video shuffle prompt.



**Question:** Measuring from the closest point of each object, what is the distance between the refrigerator and the dishwasher (in meters)?

**Ground Truth:** 1.6

<think>Let me think about this for a moment. The question is asking for the distance between the refrigerator and the dishwasher, and it specifies that the measurement should be taken from the closest point of each object. First, let's identify the refrigerator and the dishwasher in the image. The refrigerator is on the left side of the kitchen, and the dishwasher is on the right side. Now, I need to determine the closest points of these objects. The refrigerator has a handle on the left side, and the dishwasher has a handle on the right side. **The distance between these handles should give us a good estimate of the distance between the refrigerator and the dishwasher.** Now, let's estimate the distance. The refrigerator and the dishwasher are placed side by side, and there is a small gap between them. **The gap appears to be about the width of a standard kitchen appliance, which is typically around 60 cm or 0.6 meters.** Hmm, it seems like the distance between the closest points of the refrigerator and the dishwasher is approximately 0.6 meters. Oh, I see. The question is asking for the distance between the closest points, and the image provides a visual reference that supports this estimation. **Let's break it down one more time to verify. The refrigerator and the dishwasher are placed side by side, and the gap between them is about the width of a standard kitchen appliance. This confirms that the distance is indeed around 0.6 meters.** After all this reflection, I am confident that the distance between the closest points of the refrigerator and the dishwasher is approximately 0.6 meters.</think>  
 <answer>0.6</answer>

**Video-R1's reasoning**

Analyze objects

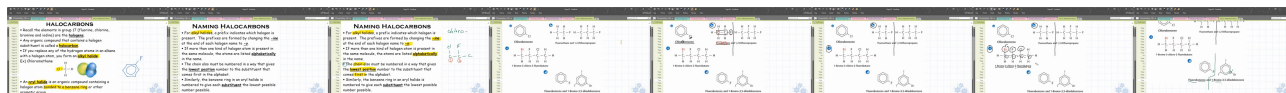
<think>The question asks for the distance between the refrigerator and the dishwasher in meters. To answer this, I need to identify the locations of these objects in the video and estimate the distance between them. Firstly, let's locate the refrigerator and the dishwasher. The refrigerator is on the left side of the kitchen, and the dishwasher is on the right side of the kitchen. **The refrigerator is a large appliance with a stainless steel door, and the dishwasher is a white appliance with a control panel at the top.** To estimate the distance, I notice the relative positions of the objects and the space around them. **The refrigerator is positioned against the wall, and the dishwasher is placed next to it. The distance appears to be a straight line between the two appliances.** The refrigerator is located near the corner of the room, and the dishwasher is slightly further away but still in the same general area. The space between them seems to be a standard kitchen layout, with no obstructions like cabinets or other appliances in the way. **The distance between the refrigerator and the dishwasher appears to be around 1.5 meters. This estimation is based on the visual scale of the room and the relative sizes of the appliances compared to the surrounding space.** Now, let's consider the other objects in the scene. The sink is in the foreground, and the stove is in the background. The window and the plants outside provide context for the setting but do not affect the measurement of the distance between the refrigerator and the dishwasher. In conclusion, the distance between the refrigerator and the dishwasher is estimated to be 1.5 meters.</think>  
 <answer>1.5</answer>

**ViSS-R1's reasoning**

Answer question

check other visual cues

Figure 14. **Additional qualitative results on VSI-Bench.** Identifying spatial locations and estimating distances or lengths between objects remain significant challenges for video reasoning models. In this example, Video-R1 estimates the distance between the refrigerator and the dishwasher primarily based on prior knowledge of typical kitchen layouts. In contrast, our ViSS-R1 model estimates the distance by referencing the proximity of nearby appliances. Furthermore, ViSS-R1 verifies its answer by cross-checking with other objects and background elements present in the video.



**Question:** For example C, if I change Br to Cl, what is the answer?

- A. 3-Chloro-2-fluoro-1-butane    B. 2-Fluoro-1,3-dichlorobutane    C. 2-Fluoro-3-chlorobutane  
 D. 1-Chloro-2-fluoro-3-chlorobutane    E. 1,3-Dichloro-2-fluorobutane

<think>Let me think about this carefully. The question is asking what the answer would be if I change Br to Cl in example C. First, let's break it down. In example C, we have a molecule with a bromine (Br) atom, a fluorine (F) atom, and a chlorine (Cl) atom. The bromine is at position 1, the fluorine is at position 2, and the chlorine is at position 3. Now, if we change the bromine (Br) to chlorine (Cl), the molecule would have a chlorine at position 1, a fluorine at position 2, and a chlorine at position 3. The numbering of the positions would remain the same, but the labels would change. So, the new molecule would be named as follows: \n- The chlorine at position 1 would be the first substituent. \n- The fluorine at position 2 would be the second substituent. \n- The chlorine at position 3 would be the third substituent. Therefore, the correct name for this molecule would be \"2-Fluoro-3-chlorobutane\". Hmm, let's verify this. The molecule has a fluorine at position 2 and a chlorine at position 3, and the chlorine at position 1 is the first substituent. This matches the structure in the options. Oh, I see. The correct answer is indeed \"2-Fluoro-3-chlorobutane\". Let's check the options:  
 A. 3-Chloro-2-fluoro-1-butane - This is incorrect because the chlorine is not at position 3.  
 B. 2-Fluoro-1,3-dichlorobutane - This is incorrect because the chlorine is not at position 1.  
 C. 2-Fluoro-3-chlorobutane - This matches our reasoning.  
 D. 1-Chloro-2-fluoro-3-chlorobutane - This is incorrect because the chlorine is not at position 1.  
 E. 1,3-Dichloro-2-fluorobutane - This is incorrect because the chlorine is not at position 1.  
 The correct answer is C.  
 Let's break it down one last time to ensure. The molecule has a fluorine at position 2 and a chlorine at position 3, and the chlorine at position 1 is the first substituent. This matches option C. /think>  
 <answer>C</answer>

Directly answer questions based on model's prior knowledge, w/o reviewing contextual cues,

**Video-R1's reasoning**

<think>Let me think about this problem step-by-step: review the naming convention provided in the first three frames  
 Firstly, I need to understand the naming convention for halocarbons. In this case, we are dealing with an alkyl halide where the halogen atoms are bonded to a carbon chain. The naming follows specific rules: the halogen is indicated by a prefix (e.g., chloro for chlorine), and the chain is numbered to give the lowest possible number to the substituent that comes first in the alphabet.  
 In the given example, the molecule is 1-Bromo-3-chloro-2-fluorobutane. If we change the bromine (Br) to chlorine (Cl), the molecule becomes 1-Chloro-3-chloro-2-fluorobutane. However, since there are two chlorine atoms, they are listed alphabetically, so the correct name is 1,3-Dichloro-2-fluorobutane.  
 Now, let's break it down further:  
 1. The chain is numbered to give the lowest possible number to the substituent that comes first in the alphabet. In this case, the fluorine atom is at position 2, the chlorine atoms are at positions 1 and 3, and the bromine atom is at position 4.  
 2. The halogen atoms are listed alphabetically. Here, the chlorine atoms come before the bromine atom.  
 If we change the bromine to chlorine, the molecule still follows the same structure but with the halogen atoms swapped. The numbering and the alphabetical order remain the same.  
 To verify the answer, I compare it with the options provided. The correct answer is 1,3-Dichloro-2-fluorobutane, which matches option E.  
 Let's consider the other options:  
 - Option A is incorrect because it changes the position of the fluorine atom.  
 - Option B is incorrect because it changes the structure and the numbering.  
 - Option C is incorrect because it changes the position of the chlorine atoms.  
 - Option D is incorrect because it changes the structure and the numbering.  
 After analyzing the options and the structure of the molecule, I conclude that the correct answer is indeed E. /think>  
 <answer>E</answer>

Analyze options

Answer question

**ViSS-R1's reasoning**

Figure 15. **Supplementary qualitative results on VideoMMU.** At the beginning of the video, detailed naming conventions for halocarbons are presented (first three frames). These important visual cues are overlooked by Video-R1, which directly attempts to answer the question based on prior knowledge, resulting in hallucinations. In contrast, our ViSS-R1 model fully utilizes the visual cues and incorporates the naming conventions, enabling it to correctly derive the answer.



**Question:** How did the cowherd use the bull's magic?

- A. Asked the magpies to form a bridge.
- B. Hurtled upwards and tried to wade through the stars.
- C. Plucked a golden hairpin and tore through the sky.
- D. Hastily placed each child in a basket.

<think>Let me think about this for a moment. The question is asking about how the cowherd used the bull's magic. Now, let's break it down and consider each option carefully.

**First, let's look at option A:** "Asked the magpies to form a bridge." This seems plausible because the cowherd did ask for help from the magpies. However, the question is specifically about the bull's magic, not the magpies.

**Next, let's consider option B:** "Hurtled upwards and tried to wade through the stars." This option describes an action taken by the cowherd, but it doesn't directly relate to the bull's magic. The bull's magic is not mentioned in this context.

**Now, let's look at option C:** "Plucked a golden hairpin and tore through the sky." This option also seems plausible. The bull's magic is mentioned in the context of the golden hairpin, which the cowherd used to tear through the sky. This action directly involves the bull's magic.

**Finally, let's consider option D:** "Hastily placed each child in a basket." This option describes an action taken by the cowherd, but it doesn't relate to the bull's magic at all.

Hmm, after considering all the options, it seems that the most accurate description of how the cowherd used the bull's magic is in option C. The bull's magic was used to pluck a golden hairpin and tear through the sky.

Oh, I see. The correct answer is C, as it directly relates to the bull's magic and the action taken by the cowherd.

Let's verify one last time. Yes, the bull's magic was used to pluck a golden hairpin and tear through the sky, which is the correct description of how the cowherd used the bull's magic.

Therefore, the correct answer is C.

Final answer: C

Let's break it down one last time to ensure. The bull's magic was used to pluck a golden hairpin and tear through the sky, which is the correct description of how the cowherd used the bull's magic.</think>

<answer>C</answer>

Directly analyze options one by one

Summarize video contents

**Video-R1's reasoning**

<think>The video is an animated sequence that tells a story involving a cowherd and a cow, set against a backdrop of a magical and fantastical world. The story begins with a scene of a woman playing a harp, followed by a landscape with a bridge and mountains. The cowherd is introduced, and he is seen interacting with the cow and the magpies. The cowherd uses the bull's magic to hurtle upwards and try to wade through the stars. He then plucks a golden hairpin and tears through the sky. The story concludes with the cowherd and the cow standing together, with the cowherd holding the golden hairpin.

Now, let's analyze the question: "How did the cowherd use the bull's magic?"

In the video, the cowherd is shown using the bull's magic to hurtle upwards and try to wade through the stars. This is depicted in a scene where the cowherd is seen flying through the sky, using the bull's magic to achieve this. The scene shows the cowherd hurtling upwards and attempting to navigate through the stars, which is a magical act performed using the bull's magic.

To answer the question, we need to identify the specific action the cowherd took with the bull's magic. In the video, the cowherd uses the bull's magic to fly and navigate through the stars. He hurtles upwards and tries to wade through the stars, which is a clear indication of how he uses the bull's magic.

Let's break it down further:

- The cowherd is seen flying through the sky.
- He uses the bull's magic to hurtle upwards.
- He attempts to wade through the stars.

The correct answer is B. Hurtled upwards and tried to wade through the stars.

The other options do not fit the scene. Option A involves the magpies forming a bridge, which is not relevant to the cowherd's use of the bull's magic. Option C involves plucking a golden hairpin and tearing through the sky, which is also not the main action. Option D involves hastily placing children in baskets, which is not related to the scene.

After analyzing the scene and the question, I can conclude my thought process and provide the answer.</think>

<answer>B</answer>

Answer specific question

Filter options

**ViSS-R1's reasoning**

Figure 16. **Supplementary qualitative results on VideoMME.** To answer the question, the Video-R1 model directly analyzes each option in isolation at the outset (highlighted in red), neglecting the rich visual context present in the colorful videos. In contrast, our ViSS-R1 model first summarizes the entire video content to form a comprehensive understanding, and then answers the question based on this holistic view. Finally, ViSS-R1 filters out incorrect options to arrive at the correct answer.