

Supplementary: Reinforcing Structured Chain-of-Thought for Video Understanding

Supplementary Material

1. EventFlow Dataset Construction

1.1. QAs Generation Pipeline

The EventFlow dataset is constructed via a three-stage automated pipeline designed to create high-quality, linguistically diverse Question-Answer (QA) pairs for training and evaluating temporal reasoning capabilities in Reinforcement Learning (RL) agents. The pipeline utilizes action annotations from existing large-scale video datasets, employs a systematic set of temporal logic templates, and leverages a Large Language Model (LLM) for linguistic diversification and question instantiation. The three core stages are:

- **Source Data Selection:** Extraction of video segments and their precise (*action, start_time, end_time*) triplets.
- **Template Design:** Definition of 15 fundamental temporal logic templates (e.g., precedence, duration, concurrency).
- **LLM Generation:** Using an LLM to rewrite templates for linguistic diversity and subsequently instantiate the questions using real action labels and derive the ground-truth answers based on temporal logic.

1.1.1. Source Datasets with Temporal Annotations

We selected three widely-used instructional and egocentric video datasets that provide dense and accurate temporal action boundary annotations. This high-quality temporal labeling is crucial for deriving objective ground-truth answers in the form of (*action, t_s, t_e*).

The composition of our dataset is as follows:

- **YouCook2:** Long instructional videos of cooking steps, contributing approximately 34% of our samples.
- **COIN:** Structured multi-step tasks, accounting for 40% of the dataset.
- **EgoExo4D:** Fine-grained "keystep" annotations from egocentric views, making up the remaining 26%.

For YouCook2 and COIN, we utilize the official training and evaluation splits. For EgoExo4D, we follow the standard keystep annotation protocols for split construction. To ensure reproducibility and eliminate LLM-induced hallucinations, the ground-truth CoT for these samples is deterministically synthesized from the human-annotated action sequences using predefined templates, rather than being generated by a large model.

1.1.2. Temporal Reasoning Templates

To cover a broad range of temporal reasoning phenomena, we design 15 template families, as summarized in Table 1. Each template represents a distinct type of temporal relation and contains placeholders (e.g., {A}, {B}, or {k}) that

are later instantiated using action labels extracted from the video annotations.

The templates are organized into a hierarchical taxonomy to evaluate two core capabilities:

Action-level temporal reasoning ($\sim 39.1\%$ of *EventFlowQA*): This category focuses on localized temporal relations around specific actions. These tasks require the model to identify immediate dependencies within a limited temporal window. For example, the *Pre-action* and *Post-action* templates (e.g., "What happened immediately before {A}?") test the model's precision in boundary localization.

Video-level temporal reasoning ($\sim 60.9\%$ of *EventFlowQA*): In contrast, these tasks require a global understanding of the entire video or long sequences. This includes *Action Ordering* (e.g., "Arrange {A, B, C} in order"), *Action Counting*, and *Long-term Dependency* reasoning. These questions are specifically designed to challenge the model's ability to maintain a coherent *Summary* and *Think* process without drift over extended horizons.

As shown in Table 1, this hierarchical design spans across diverse reasoning types including Sequence, Duration, and Frequency. By mapping human-annotated action boundaries to these templates, we ensure that the resulting ground-truth CoT paths are both factually grounded and structurally rigorous, providing a robust signal for our single-stage RL training.

1.1.3. LLM-Driven Question Generation

Template Rewriting using an LLM To mitigate the risk of RL models overfitting to simple, repetitive sentence structures, we utilized a state-of-the-art LLM to increase the linguistic diversity of our templates.

- **Process:** Each of the 15 core templates was fed to the LLM with a prompt instructing it to generate multiple paraphrased variations (typically 10-15 per template).
- **Constraint:** The LLM was strictly constrained to maintain the exact temporal logic and the placement of the action placeholders, ensuring that the rewritten query remained logically identical to the core template.
- **Result:** This process expanded the template pool from 15 to hundreds of unique question structures, significantly enhancing the natural language robustness of the dataset.

Action Grounding and Slot Filling The final stage involves generating the executable QA pairs by merging the rewritten templates with the video data and deriving the ground-truth answer.

Table 1. Overview of the temporal reasoning templates used in the EventFlow dataset, grouped into action-level and video-level categories.

Category	Description	Example Template
Action-Level Temporal Reasoning		
Action Order Reasoning	Identify the action that happens immediately before or after a given action.	“What action happened immediately before {A}?”
Temporal Causality	Select the action that causes or results from a given action.	“What was the most recent action that led to {A}?”
Action Anticipation	Predict the most likely next action following a given step.	“What action will the person perform immediately after {A}?”
Sequential Prediction	Choose the next likely action(s) in an ongoing sequence.	“Predict the next {k} actions after {A}.”
Duration Estimation	Estimate how long an action takes.	“How long does the action {A} take?”
Temporal Localization	Find the exact time interval when an action happens.	“When does the action {A} occur in the video?”
Temporal QA from Narration	Choose the action that aligns with a narrated timestamp.	“What action is taking place at time ratio {t}?”
Video-Level Temporal Reasoning		
Temporal Gap Estimation	Estimate the time gap between two actions.	“How much time passed between {A} and {B}?”
Temporal Yes/No	Answer whether one action happened before another.	“Did {A} happen before {B}?”
Temporal Comparison	Choose which action occurred earlier or later.	“Which action occurred first?”
Duration Comparison	Choose the action that took more or less time.	“Which action took the longest?”
In-between Action	Identify actions that happened between two given actions.	“Which actions occurred between {A} and {B}?”
Transcription	Identify all actions in correct order.	“Which option best represents the sequence of actions in the video?”
Action Order Reasoning (Extreme)	Select the first or last action from a sequence.	“Which action happened first in the video?”
Action Order Sorting	Sort multiple actions in the correct chronological order.	“Arrange these actions in the correct order: {A,B,C}.”
Order Consistency	Identify which option maintains correct chronological sequence.	“Which option shows the actions in the correct chronological order?”

- **Selection and Insertion:** A randomly selected rewritten template is paired with relevant action segments extracted from the source video datasets. The action labels (e.g., “chop vegetables,” “heat oil”) are inserted into the placeholders.
- **Ground Truth Derivation:** The ground-truth answer is programmatically computed solely based on the temporal logic derived from the video’s t_s and t_e annotations. For instance, if a query is of the “Precedes” type, the answer is determined by verifying the condition $t_e(A) < t_s(B)$ against all possible actions A within the context. This rigorous, data-driven approach ensures the absolute correctness of every generated answer.

1.2. Output Format and Examples

As shown in Listing 1, a sample instance of EventFlow includes the following key information required for training and evaluation:

- **Video ID:** Unique identifier of the source video.
- **Video interval:** The specific temporal span used as context.
- **Question:** The instantiated, natural language question.
- **Answer:** The programmatically derived ground-truth action label, time value, or Boolean response.
- **Query type:** The core temporal logic category (e.g., Sequential, Durational).

Listing 1. Example JSON entry from our constructed dataset.

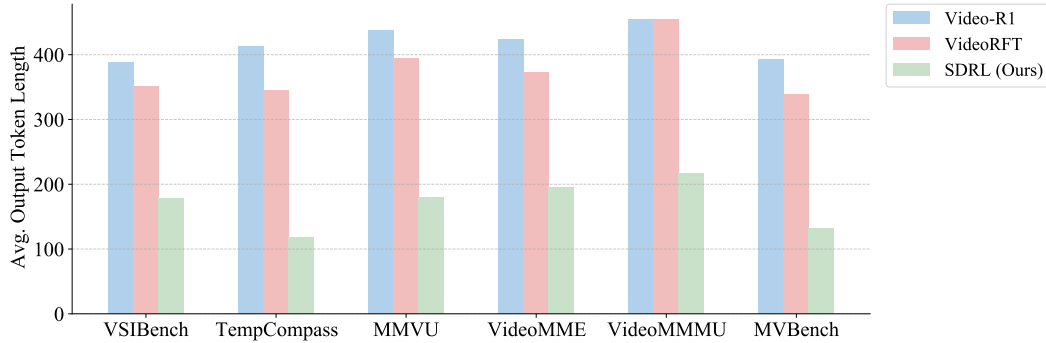


Figure 1. Comparison of the average output token length across six video benchmarks. We observe that Video-R1 consistently produces longer reasoning sequences than VideoRFT and SDRL, while SDRL yields the most concise outputs across all datasets.

```
{
  "tag": "action level",
  "type": "temporal_neighbor",
  "template_type": "MCQ",
  "answer_type": "choice",
  "category": "Action Order Reasoning",
  "template": "What action happened immediately before {} in the video?",
  "new_template": "What activity took place directly prior to {} in the video?",
  "question_args": [
    "'wipe screen again'"
  ],
  "choices": "A. paste protector on the screen\nB. remove the label\nC. place label\nD. wipe screen",
  "final_answer": "B",
  "answer_text": "remove the label",
  "video_interval": [32.39, 134.19],
  "act_idx": 4
}
```

1.3. Data Distribution Analysis

To provide a granular understanding of EventFlowQA, we analyze the distribution of question types, as illustrated in Fig. 2. The dataset comprises a 50K training split and a 3K evaluation split. It is structured into two hierarchical levels of temporal reasoning: *action-level* and *video-level*. Action-level questions account for 39.1% of the dataset, focusing on localized reasoning over specific actions or short temporal segments. In contrast, video-level questions constitute the majority (60.9%), requiring a holistic understanding of long-range temporal contexts and inter-event relationships.

The fine-grained distribution across 15 temporal reasoning categories is depicted in the outer ring of Fig. 2. These categories span a diverse range of skills, including action ordering, duration estimation, and causal reasoning. Notably, *Action Order Reasoning* (11.5%), *Temporal Localization* (8.7%), and *Duration Estimation* (8.5%) are the most

prominent categories, reflecting the dataset’s emphasis on capturing the underlying temporal structure of videos.

Furthermore, challenging scenarios such as *Sequential Prediction*, *Action Anticipation*, and *In-between Action* are included to evaluate a model’s ability to infer implicit temporal relations. This multi-faceted distribution ensures that EventFlowQA serves as a comprehensive and demanding benchmark for assessing the temporal reasoning capabilities of modern video-language models.

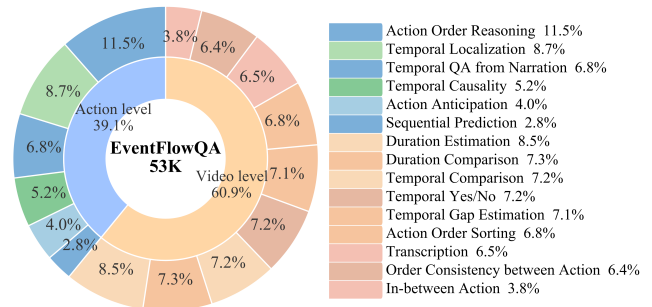


Figure 2. Data distribution of the EventFlowQA-53K dataset. The dataset consists of two hierarchical levels: Action-level questions and Video-level questions. The outer ring shows the detailed proportions of each temporal reasoning category, including action order, duration, comparison, causality, anticipation, and others.

2. Prompt For Training and Inference

Our prompt used for both training and inference is shown in Table 2. It introduces a summary tag to encourage the model to summarize the video before generating its reasoning and final answer.

3. Different Tags for Summary

To encourage the model to produce temporally ordered summaries, we explore a series of tags that explicitly guide the model’s initial summarization behavior, shown in Ta-

Table 2. Prompt for training and inference.

Prompt Template	
<p>{Question} {Type of Template} Given the video clip and the question, first summarize the sequence of actions observed in the video using <summary></summary>tags. Then, use that summary to guide your reasoning inside <think></think>tags. Finally, state the answer clearly in <answer></answer>tags.</p>	
Type of Template	
multiple choice	Please provide only the single option letter (e.g., A, B, C, D, etc.).
numerical	Please provide the numerical value (e.g., 42 or 3.14).
OCR	Please transcribe text from the image/video clearly and provide your text answer.
free-form	Please provide your text answer.
regression	Please provide the numerical value (e.g., 42 or 3.14).

ble 3. Our objective is to find a tag that naturally induces the model to output the sequence of actions in the correct chronological order, even before applying any reinforcement learning. In other words, a suitable tag should enable the model to produce summaries with high initial accuracy and clear temporal structure.

We experiment with multiple instruction tags, each designed to prompt the model toward a different style of temporal decomposition, for example, requesting the model to “list the sequences of **actions**,” “describe the sequence of **events**” or “**summarize** the sequence of actions.” Although these tags all aim at temporal organization, they lead to substantially different model behaviors.

For each tag, we compute the initial accuracy. This quantitative comparison on Table 3 allows us to assess which tag most effectively supports temporal grounding prior to RL optimization. To further understand the qualitative differences introduced by each tag, we visualize representative summaries generated under different tag settings. Based on this combined quantitative and qualitative analysis, we select the tag that achieves the highest initial accuracy. This tag subsequently serves as the default instruction during structured summary generation in our framework.

4. Efficiency Analysis

Analysis of Output Token Length. Our method, SDRL, driven by auxiliary summaries and optimized via Reinforcement Learning (RL), exhibits a clear and beneficial characteristic in its generated output (CoT/Summary): a **significantly shorter average token length** compared to baseline MLLMs (Video-R1 and VideoRFT), as demonstrated in Figure 1. This statistical finding is crucial for interpreting the model’s behavior and the effectiveness of our optimization strategy in complex video understanding tasks.

- **Efficiency and Conciseness** As shown in Figure 1, the average output length of SDRL is consistently the shortest across all six video benchmarks. This strong indication of conciseness demonstrates the model’s overall **high efficiency** in generating explanations.
- **Successful Consistency in Summary** The RL framework, guided by the summary, effectively regularizes the model’s output generation process. This optimization encourages the MLLM to prune redundant or repetitive reasoning steps, focusing only on the minimum essential information required for accurate prediction.
- **High-Quality Outputs** The reduced token count is not merely compression, coupled with our superior performance, this short length confirms a successful optimization towards **high-quality outputs**. This characteristic directly translates to practical advantages, including reduced inference latency and lower computational overhead during deployment.

In summary, the substantial reduction in average output length confirms that SDRL learns a highly effective and concise reasoning path, demonstrating that our RL optimization successfully yields a model that is both powerful and efficient.

5. Stability and Factual Grounding of CVK

A key concern regarding self-supervised consistency is whether enforcing agreement among sampled summaries might lead to “mode collapse” or the reinforcement of hallucinations. We provide empirical evidence to show that our CVK objective remains stable and factually grounded throughout training.

Mechanism of Stability. As defined in Eq. 8 in the main paper, the consistency anchor is derived exclusively from the group members that yield the *correct* final answer. This filtering mechanism ensures that the model aligns towards reasoning paths that are not only consistent but also functionally effective, preventing the mutual reinforcement of erroneous summaries during early training stages.

Quantitative Factuality. As shown in Table 4 in the main paper, the self-supervised CVK significantly enhances summary quality, achieving a +3.37% improvement in BLEU and +9.43% in sBERT compared to the base-

Table 3. Effect of different prompt tags on model behavior. For each tag, the left column shows the question prompt and the right column shows the model output. Accuracy values reflect the QA performance under each tag. The examples show how prompt structure shapes the granularity, ordering, and coherence of the generated summaries.

Question Prompt (tag, accuracy)	Model Output
 <p>action (26.45%)</p> <p>Given a video clip, first list the sequence of actions observed in the video in order in <action></action>tags. Then, answer the question with output the thinking process in <think></think>tags, followed by the final answer in <answer></answer>tags.</p>	<p><action>1. put 2 tsp of olive oil add cut bread dices and fry 2. tear fresh lettuce leaves and add them with seasoned kale leaves 3. cook bacon in a pan until crispy 4. add roasted bread crumbs crispy bacon slices of parmesan cheese shaved to seasoned greens and pour dressing blend on top</action></p>
<p>caption (26.95%)</p> <p>Given a video clip, first describe the sequence of actions observed in the video in the exact order they occur using <caption></caption>tags. Then, answer the question with output the thinking process in <think></think>tags, followed by the final answer in <answer></answer>tags.</p>	<p><caption> A woman is standing in her kitchen. She picks up a piece of bread and places it on a cutting board. She then cuts the bread into slices. Next, she puts some butter in a frying pan and cooks the foie gras. After that, she cooks the apple slices in the same pan. Finally, she places the apple slices and the foie gras on a plate.</caption ></p>
<p>event (16.12%)</p> <p>Given a video clip, first describe the sequence of events observed in the video in the exact order they occur using <event></event>tags. Then, answer the question with output the thinking process in <think></think>tags, followed by the final answer in <answer></answer>tags.</p>	<p><event> cut the foie gras into slices </event> <event> pour the juice onto the foie gras</event> <event> place the apples and the foie gras on a plate</event> <event> sprinkle salt and pepper on top</event></p>
<p>segment (23.53%)</p> <p>Given a video clip, first summarize the sequence of actions observed in the video using <segment></segment>tags. Then, output the thinking process in <think></think>tags, followed by the final answer in <answer></answer>tags.</p>	<p><segment>0:12.3 - 0:14.6</segment> <segment>0:15.0 - 0:17.8</segment> <segment>0:19.0 - 0:20.2</segment> <segment>0:20.6 - 0:21.7</segment> </p>
<p>summary (30.96%)</p> <p>Given a video clip, first summarize the sequence of actions observed in the video using <summary></summary>tags. Then, answer the question with output the thinking process in <think></think>tags, followed by the final answer in <answer></answer>tags.</p>	<p><summary> A woman is preparing a dish in her kitchen. She starts by cutting a piece of bread and placing it on a plate. Then, she cuts an apple and places it on the same plate. Next, she cooks the bread in a pan with butter. After that, she cooks the apple slices in another pan. Finally, she places the cooked bread and apple slices on a plate. </summary></p>

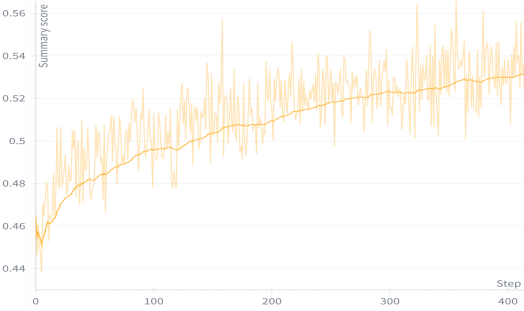


Figure 3. Summary quality during training measured by $(\text{BLEU} + \text{sBERT})/2$. The curve shows steady improvement without collapse.

line. These gains indicate that enforcing consistency among “correct” paths naturally encourages the model to capture more accurate semantic and temporal information from the video.

Training Dynamics. We monitor the summary quality score, $(\text{BLEU} + \text{sBERT})/2$, across the training process. As illustrated in Fig. 3, the quality score improves monotonically and stabilizes as the model converges. There is no evidence of performance degradation or collapse, confirming that CVK effectively regularizes the reasoning space without sacrificing factual integrity. Qualitative examples in Fig. 6 further demonstrate that SDRL produces more temporally-precise summaries than previous RL-based methods.

6. More Visualization

We provide qualitative visualizations across several datasets to illustrate the behavior of different reasoning strategies under various video understanding scenarios. Figures 4, 5, 6, and 7 show examples from different benchmarks. Each visualization includes the input video frames, the corresponding question and answer choices, and the generated reasoning outputs from Video-R1, Video-RFT, and our SDRL. Across diverse reasoning tasks, SDRL consistently produces concise and well-structured summaries that focus on the most salient visual cues necessary for the task. These visualizations highlight how SDRL’s summary-driven reasoning encourages grounded, task-relevant interpretation of the video, providing clearer and more consistent reasoning trajectories compared to the other baselines.

7. Failure Case

To illustrate the limitations of summary-driven reasoning without explicit ground-truth supervision, we analyze a representative failure case shown in Figure 8. Although the video contains a long sequence of visually similar cutting actions, the SDRL model generates a summary that

omits several intermediate steps and mistakenly focuses on a salient but non-final “spooning” action. This incomplete and temporally misaligned summary propagates into the reasoning process, leading the model to incorrectly select option C, despite the true final action being “cut off top and chop into blocks.” The example highlights a core challenge of unsupervised summary generation: when low-level visual patterns are similar across actions, the model may gravitate toward semantically plausible but temporally incorrect interpretations, ultimately harming downstream question answering.

Moreover, as shown in Figure 9, when the video involves fast and fine-grained object manipulations, SDRL tends to produce a coarse, high-level description rather than enumerating each atomic action. In this example of an occlusion game, the model summarizes the sequence as “moving bowls around,” without tracking the precise order of swaps and lifts that determine the final hidden location. This loss of fine-grained temporal detail causes the reasoning chain to rely on a generic interpretation of the game instead of the actual step-by-step motion, ultimately leading SDRL to infer an incorrect hiding position. The case demonstrates that, without explicit supervision on action boundaries, SDRL summaries may abstract away critical micro-actions, resulting in temporally drifting or semantically incomplete reasoning.

8. Limitations and Future Work

Despite the effectiveness of our SDRL framework in improving temporal reasoning, it still exhibits certain limitations. First, in scenarios involving long-horizon videos with dense or fine-grained actions, the model may generate relatively coarse summaries that omit micro-level temporal cues. This abstraction bias can lead to temporal drift when tasks demand precise frame-level tracking. Second, while CVK enforces consistency among high-quality reasoning paths, it does not provide a formal guarantee of absolute factual correctness. In cases where answer supervision is sparse, there remains a theoretical risk that the model might converge on consistent but partially hallucinated summaries. Furthermore, our single-stage RL framework relies on the base model’s inherent instruction-following capabilities to parse structured tags; adapting SDRL to significantly weaker models without such priors may require a brief warm-up phase.

Future work will explore incorporating stronger temporal supervision, such as contrastive action boundaries and structured sequence alignment, to encourage finer-grained summary generation. Additionally, we plan to integrate external verification signals or human-in-the-loop evaluation to further anchor the factuality of the reasoning process and investigate multi-step self-refinement strategies to mitigate summary drift in densely interactive video scenarios.

Video:



Question:

Which description is a more suitable match for the video?

- A. The basketball is moving from right to left.
- B. The basketball is moving from left to right.

Video-R1:

<think> Let me think about this for a moment. The question is asking which description is a more suitable match for the video, and we have two options to choose from. First, let's break it down and analyze the video. The basketball is clearly in motion, and we need to determine its direction. The player is on the right side of the frame, and the basketball is moving towards the hoop, which is on the left side of the frame. Now, let's consider the options: A. The basketball is moving from right to left. B. The basketball is moving from left to right. Hmm, it seems like the basketball is indeed moving from the right side of the frame towards the left side. The player is on the right, and the basketball is heading towards the hoop, which is on the left. This suggests that the basketball is moving from right to left. Wait, let's double-check. If the basketball were moving from left to right, it would be moving away from the hoop, which is not the case. The basketball is clearly heading towards the hoop, which is on the left side of the frame. Oh, I see. The basketball is indeed moving from right to left. This aligns with option A. Let's verify one last time. The player is on the right, and the basketball is moving towards the left, which is the hoop. This confirms that the basketball is moving from right to left. After all this reasoning, I am confident that the more suitable match for the video is option A. Let's break it down one last time to ensure. The basketball is moving from the right side of the frame towards the left side, which is the hoop. This is a clear indication that the basketball is moving from right to left. Final answer: A Let's verify one last time. The basketball is moving from right to left, and this matches option A perfectly.</think>

<answer>B</answer> ❌

Video-RFT:

<think>To determine the correct description, let's analyze the video step-by-step. The video begins with a silhouette of a person preparing to shoot a basketball at sunset. The person then jumps and releases the ball, which is captured mid-air. The ball is shown moving towards the hoop, and the video concludes with the ball in the hoop. The key elements are the direction of the ball's movement. The ball starts on the left side of the frame and moves towards the hoop, which is on the right side. This indicates that the ball is moving from left to right. Therefore, the correct description is option B, which states that the basketball is moving from left to right. Option A suggests that the basketball is moving from right to left, which contradicts the visual evidence. The ball starts on the left and moves towards the right, confirming that the correct answer is B. To further confirm, let's consider the sequence of events. The person prepares to shoot, jumps, and releases the ball. The ball is captured mid-air, moving towards the hoop. The hoop is on the right side, and the ball is moving towards it. This visual evidence clearly shows that the ball is moving from left to right. Therefore, the correct answer is B. In summary, the video shows the basketball moving from left to right towards the hoop. The person is on the left side of the frame, and the ball is moving towards the right side. The correct description is option B, which states that the basketball is moving from left to right. The other option, A, is incorrect because it suggests the ball is moving from right to left, which is not the case. Thus, the correct answer is B. </think>

<answer>B</answer> ❌

SDRL:

<summary> The video shows a basketball player shooting a ball towards a hoop. The ball is seen in mid-air, heading towards the hoop. The player is on the right side of the frame, and the hoop is on the left. The sky is colorful with a sunset, and the ball is moving from the player towards the hoop.

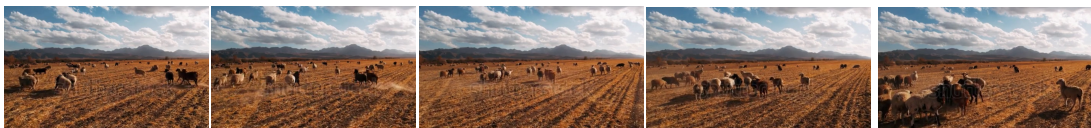
</summary>

<think> The basketball is moving from the player towards the hoop, which is on the left side of the frame. This indicates that the ball is moving from right to left relative to the player's position. </think>

<answer>A</answer> ✅

Figure 4. An example on Temcompass dataset. Video-R1 and Video-RFT generate lower information density descriptions (All red part is the video description). Video-R1's description is iterative, which may contribute to a longer reasoning process. This extended path may be associated with the observed instability in the inference chain (yellow highlighting), where intermediate thoughts contradict the final outcome, resulting in the selection of the incorrect answer. Video-RFT's context is more dispersed, which potentially leads to an initial misrepresentation of the visual facts. SDRL produces a concise, highly focused summary by directly extracting the necessary visual facts ("ball on the right, hoop on the left"). This concentrated context supports a shorter and more direct reasoning path.

Video:



Question:

Which sentence better captures the essence of the video?

- A. The sheep is eating grass.
- B. A sheep is running.

Video-R1:

<think>Let me think about this for a moment. The question is asking which sentence better captures the essence of the video. To answer this, I need to carefully analyze the content of the video and the options provided. First, let's break down the video. **The video shows a large flock of sheep grazing in a field with mountains in the background. The sheep are scattered across the field, and some are standing still while others are moving around. The overall scene is peaceful and serene, with the sheep engaged in their natural behavior of grazing.** Now, let's consider the options: **A. The sheep is eating grass.** - This sentence is true, but it only describes one specific action of the sheep. While it is accurate, it doesn't capture the essence of the entire scene. The video shows multiple sheep, and not just one, all engaged in the same activity. **B. A sheep is running.** - This sentence is also true, but it focuses on a single action of one sheep. While there is a sheep running in the video, it is not the main focus. **The video is more about the overall scene of the sheep grazing in the field.** Hmm, it seems that neither option fully captures the essence of the video. However, if I have to choose the better option, it would be the one that provides a broader description of the scene. **The video is more about the sheep grazing in a field,** which is a more encompassing description than focusing on a single action. Oh, I see. The better sentence that captures the essence of the video is the one that describes the overall scene of the sheep grazing in the field. Therefore, the correct answer is the one that includes the broader description. Let's verify this one last time. The video is indeed more about the sheep grazing in the field, and the sentence \"The sheep is eating grass\" is a more accurate and encompassing description than \"A sheep is running.\" Final answer: A Let's break it down one last time to ensure. **The video is about the sheep grazing,** and the sentence \"The sheep is eating grass\" is the better capture of the essence.</think>

<answer>A</answer> ❌

Video-RFT:

<think>To determine which sentence better captures the essence of the video, let's analyze the content step-by-step. **The video shows a flock of sheep grazing in a vast, open field with golden-brown grass and a mountainous background. The sheep are scattered across the field, some standing still while others move around, creating a dynamic scene. The sky is partly cloudy, adding to the serene atmosphere.** The key elements are the sheep, the field, and the mountainous background. The sheep are the main subjects, and their actions are central to the video. The field and mountains provide the setting, but the sheep's behavior is the primary focus. **Option A states,** \"The sheep is eating grass.\" This sentence is singular and describes a single sheep's action. However, **the video shows a flock of sheep, not just one.** The sheep are grazing, which involves eating grass, but the sentence is too specific and doesn't capture the overall scene. **Option B says,** \"A sheep is running.\" This sentence is also singular and describes a single sheep's action. **While the video does show some sheep moving around, the overall scene is more about the flock grazing.** The sentence is too narrow and doesn't encompass the entire video. The video's essence is about the flock of sheep grazing in a beautiful, open field with mountains in the background. The sheep's actions are part of this broader scene. Therefore, the best option is A, as it captures the main activity of the sheep eating grass, which is a key element of the video. The other options are too specific and don't encompass the entire scene. In conclusion, the best option is A, as it captures the main activity of the sheep eating grass, which is a key element of the video. </think>

<answer>A</answer> ❌

SDRL:

<summary> **The video shows a flock of sheep grazing in a field with mountains in the background. The sheep are scattered across the field, some standing still while others are moving around. The sky is partly cloudy, and the overall scene is peaceful and serene.** </summary>

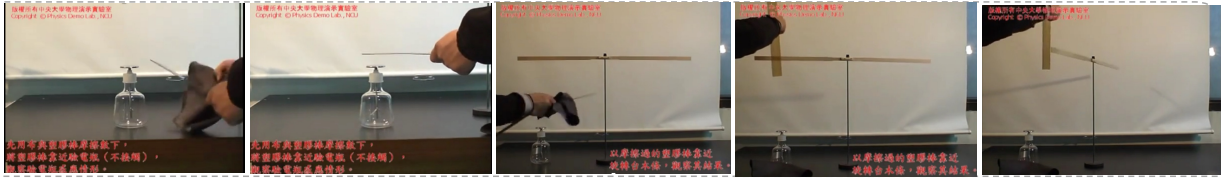
<think> **The video primarily depicts a flock of sheep in a field, with some sheep running and others grazing.** The question asks for the sentence that better captures the essence of the video. **Option A,** \"The sheep is eating grass,\" is more general and **doesn't capture the movement** of the sheep. **Option B,** \"A sheep is running,\" is more specific and **captures the action** of one sheep in the video.

</think>

<answer> B </answer> ✅

Figure 5. An example on Temcompass dataset. Although all three methods describe roughly the same scene (descriptions in red), their narrative focus differs significantly. SDRL produces a concise and coherent summary that preserves global context, enabling the model to capture the overall movement pattern of the sheep (green highlighting) and correctly identify the key action in the video. In contrast, Video-R1 and Video-RFT provide fragmented or overly broad descriptions that dilute the salient motion cues, causing the reasoning to focus on partial details and miss the essential dynamics of the scene.

Video:



Question:

Why does the wooden stick rotate with the plastic strip?

- A. Wooden strips and plastic rods with different electrical charges attract each other.
- B. The wooden stick and the plastic rod have the same electric charge and repel each other.",
- C. The plastic rod is charged, but the wooden strip is uncharged. The plastic rod ionizes the air to make the wooden strip charged and creates a force interaction with the plastic rod.
- D. The wooden strips are charged, but the plastic rods are not. The molecules of the plastic rod are polarized and the wooden strips are subject to magnetic forces.
- E. The wooden stick is uncharged and the plastic rod is charged. The wood molecules are polarized and rotate due to uneven forces in the electric field.

Video-R1:

<think>Let me think about this for a moment. The question is asking why the wooden stick rotates with the plastic strip. The options provide various reasons, and I need to carefully consider each one. First, let's break down the options: A. Wooden strips and plastic rods with different electrical charges attract each other. - This seems plausible because opposite charges attract. However, the question doesn't specify the charges, so this might not be the most accurate. B. The wooden stick and the plastic rod have the same electric charge and repel each other. - This is less likely because repulsion would cause the stick to move away, not rotate. C. The plastic rod is charged, but the wooden strip is uncharged. The plastic rod ionizes the air to make the wooden strip charged and creates a force interaction with the plastic rod. - This seems quite plausible. The plastic rod being charged can ionize the air, making the wooden strip charged and causing it to rotate. D. The wooden strips are charged, but the plastic rods are not. The molecules of the plastic rod are polarized and the wooden strips are subject to magnetic forces. - This is less likely because there's no mention of magnetic forces in the question. E. The wooden stick is uncharged and the plastic rod is charged. The wood molecules are polarized and rotate due to uneven forces in the electric field. This is similar to option C but focuses more on the polarization of the wood molecules. Hmm, after considering all the options, C seems the most logical. The plastic rod being charged and ionizing the air to make the wooden strip charged is a common explanation for such phenomena. Oh, I see. The correct answer is C because it provides a clear and plausible explanation for the rotation of the wooden stick. Let's verify one last time. Yes, C fits the scenario best.</think>

<answer>C</answer> ❌

Video-RFT:

<think>To determine why the wooden stick rotates with the plastic strip, let's review the video content. The video begins with the National Central University Physics Demonstration Lab logo, followed by a demonstration involving a glass bottle and a plastic strip. The person rubs the plastic strip with a cloth, generating static electricity. The bottle is then brought near the strip, and the stick attached to the bottle rotates, indicating an electrostatic interaction. The key observation is that the stick rotates when the bottle is brought near the strip. This rotation suggests that the plastic strip, which has been rubbed and thus charged, is attracting or repelling the stick. The stick's rotation implies that it is being pulled towards or pushed away from the charged strip. Let's evaluate the given options: - Option A states that the wooden strips and plastic rods have different electrical charges, which could lead to attraction or repulsion. - Option B suggests that both have the same charge, leading to repulsion, but the video shows the stick rotating, indicating attraction. - Option C proposes that the plastic rod is charged and ionizes the air, making the stick charged and creating a force interaction. This aligns with the observed rotation. - Option D mentions magnetic forces, which are not relevant to static electricity. - Option E suggests that the stick is uncharged and the plastic rod is charged, leading to polarization and rotation. However, the video shows the stick rotating, indicating that it is charged. Given the video's demonstration, the most plausible explanation is that the plastic rod is charged and ionizes the air, making the stick charged and creating a force interaction. Therefore, the correct answer is C. In summary, the video shows that the plastic strip, after being rubbed, is charged and ionizes the air, making the stick charged and creating a force interaction. This results in the stick rotating when the bottle is brought near the strip.</think>

<answer>C</answer> ❌

SDRL:

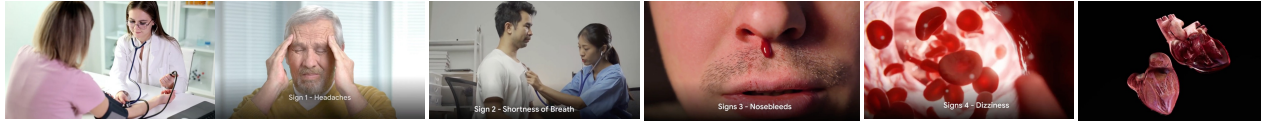
<summary> The video shows a person using a plastic strip to charge a glass bottle. The bottle is then brought close to a wooden stick mounted on a stand. The stick rotates when the bottle is brought near it. The person then uses the charged plastic strip to bring it close to the stick, causing it to rotate further.</summary>

<think> The rotation of the wooden stick when the charged plastic strip is brought near it suggests an interaction due to electric charges. The plastic strip, when rubbed, likely becomes charged, and this charge interacts with the stick, causing it to rotate. The stick's rotation indicates that it is also charged, as like charges repel and opposite charges attract. The video does not provide enough detail to determine the exact nature of the charges (positive or negative), but the rotation implies a force interaction.</think>

<answer> E </answer> ✅

Figure 6. An example on MMVU dataset. Video-R1 provides almost no grounding in the actual video content and instead jumps directly into option analysis, making it highly susceptible to thinking drift. Video-RFT includes an initial video description, but its reasoning process gradually shifts away from the visual evidence, leading to potential drift as well. SDRL maintains tight grounding by consistently referencing key visual cues throughout the reasoning process, thereby preventing drift and enabling more reliable decision-making.

Video:



Question:

Which of the following pathophysiological mechanisms is most likely being demonstrated, and which biomedical engineering intervention could most effectively manage this condition?

- A. Atherosclerosis leading to myocardial infarction, managed by coronary stent implantation.
- B. Chronic hypertension causing intracerebral hemorrhage, managed by implantable blood pressure sensors for real-time monitoring.
- C. Hemophilia causing spontaneous bleeding, managed through recombinant clotting factor therapy.
- D. Sickle-cell disease causing vascular occlusions, managed by gene therapy to correct hemoglobin mutations.
- E. Deep vein thrombosis leading to pulmonary embolism, managed by developing anticoagulant drug delivery systems.

Video-R1:

<think>Let me think about this carefully. The question is asking about the pathophysiological mechanism being demonstrated and the most effective biomedical engineering intervention for managing it. The images and the context provided seem to revolve around cardiovascular and cerebrovascular issues, which narrows down our options significantly. First, let's break it down: - The images show a heart, blood vessels, and a brain, which are all related to the cardiovascular and cerebrovascular systems. - The mention of "Signs 3 - Nosebleeds" and "Signs 4 - Dizziness" suggests that the focus is on symptoms related to these systems. Now, let's consider each option: A. Atherosclerosis leading to myocardial infarction, managed by coronary stent implantation. - This option involves atherosclerosis, which is a common cause of coronary artery disease and can lead to myocardial infarction. The intervention of coronary stent implantation is a well-known treatment for this condition. B. Chronic hypertension causing intracerebral hemorrhage, managed by implantable blood pressure sensors for real-time monitoring. - This option involves hypertension, which can lead to intracerebral hemorrhage. The intervention of implantable blood pressure sensors is a plausible approach for managing hypertension. C. Hemophilia causing spontaneous bleeding, managed through recombinant clotting factor therapy. - This option involves hemophilia, which is a bleeding disorder. The intervention of recombinant clotting factor therapy is a well-known treatment for this condition. D. Sickle-cell disease causing vascular occlusions, managed by gene therapy to correct hemoglobin mutations. - This option involves sickle-cell disease, which is a genetic disorder. The intervention of gene therapy is a promising approach for managing this condition. E. Deep vein thrombosis leading to pulmonary embolism, managed by developing anticoagulant drug delivery systems. - This option involves deep vein thrombosis, which can lead to pulmonary embolism. The intervention of anticoagulant drug delivery systems is a plausible approach for managing this condition. Hmm, after considering all the options, it seems that the most relevant and effective intervention for the pathophysiological mechanism demonstrated in the images is related to atherosclerosis and myocardial infarction. Oh, I see. The most likely pathophysiological mechanism being demonstrated is atherosclerosis leading to myocardial infarction, and the most effective biomedical engineering intervention is coronary stent implantation.</think>

<answer>A</answer> ❌

Video-RFT:

<think>To determine the correct pathophysiological mechanism and corresponding biomedical engineering intervention, let's review the video content step-by-step. The video begins with a doctor examining a patient, followed by a close-up of a nosebleed, indicating signs of nosebleeds. It then transitions to a blood clot in the brain, suggesting signs of dizziness. The video also shows a heart with blood flowing through it, indicating signs of heart issues. Finally, it depicts a foot with a blood clot, indicating signs of blood clots. The video's sequence of signs nosebleeds, dizziness, heart issues, and blood clots points to a condition involving blood clotting and vascular issues. Let's evaluate the given options: Option A: Atherosclerosis leading to myocardial infarction, managed by coronary stent implantation. This option focuses on heart issues, which are not the primary focus of the video. The video emphasizes blood clots and vascular issues, not atherosclerosis. Option B: Chronic hypertension causing intracerebral hemorrhage, managed by implantable blood pressure sensors for real-time monitoring. This option addresses high blood pressure leading to brain bleeding, which is not the main focus of the video. The video emphasizes blood clots and vascular issues, not hypertension. Option C: Hemophilia causing spontaneous bleeding, managed through recombinant clotting factor therapy. This option addresses bleeding disorders, which are not the primary focus of the video. The video emphasizes blood clots and vascular issues, not bleeding disorders. Option D: Sickle-cell disease causing vascular occlusions, managed by gene therapy to correct hemoglobin mutations. This option addresses vascular occlusions due to sickle-cell disease, which aligns with the video's focus on blood clots and vascular issues. The video's sequence of signs nosebleeds, dizziness, heart issues, and blood clots suggests a condition involving blood clots and vascular issues, which is consistent with sickle-cell disease. Option E: Deep vein thrombosis leading to pulmonary embolism, managed by developing anticoagulant drug delivery systems. This option addresses blood clots in the legs leading to pulmonary embolism, which is not the primary focus of the video. The video emphasizes blood clots and vascular issues, not deep vein thrombosis. Based on the video's focus on blood clots and vascular issues, the correct answer is D: Sickle-cell disease causing vascular occlusions, managed by gene therapy to correct hemoglobin mutations.</think>

<answer>D</answer> ❌

SDRL:

<summary> The video begins with a doctor examining a patient, possibly checking for signs of illness. It then transitions to a close-up of a blood vessel, followed by a scene of a blood sample being examined under a microscope. Next, a nurse is shown taking a patient's blood pressure. The video then cuts to a close-up of a heart, and finally, a scene of a patient being treated with a medical device. </summary>

<think> The sequence of actions in the video suggests a focus on medical examination and treatment, particularly related to blood and heart health. The examination of the blood vessel and the blood sample under a microscope, along with the blood pressure check and the heart imagery, point towards cardiovascular health and potential blood-related issues. The treatment scene with the medical device could imply a biomedical engineering intervention.</think>

<answer> B </answer> ✅

Figure 7. An example on MMVU dataset. Both Video-R1 and VideoRFT devote excessive attention to analyzing the answer options, producing overly long reasoning chains that deviate from the actual visual evidence and easily lead to thinking drift. In contrast, SDRL anchors the reasoning in the video content, maintains a concise and focused explanation, and thus avoids unnecessary detours caused by option-driven overanalysis.

Video:



Question:

What is the last action performed in the video?

- A. cut along back bone and remove loin
- B. cut off top and chop into blocks
- C. spoon away the leftover meat
- D. cut one loin into three pieces

Segment GT: The video shows the actions in order: chop off head of tuna, cut down middle, cut along back bone and remove loin, spoon away the leftover meat, cut one loin into three pieces, cut away bloodline, cut off top and chop into blocks

SDRL:

<summary> The video shows a person preparing a large fish by cutting it open, removing the backbone, and then cutting the meat into smaller pieces. The final frame shows the person using a spoon to remove any remaining meat from the fish. </summary>

<summary>

<think> The sequence of actions includes: 1. Opening the fish and removing the backbone. 2. Cutting the fish into smaller pieces. 3. Using a spoon to remove any leftover meat. The last action described in the video is the use of a spoon to remove the leftover meat, which aligns with option C. </think>

<answer> C </answer> ❌

Figure 8. A failure case from EventFlow dataset. Without ground-truth supervision, SDRL may produce incomplete or temporally mis-aligned summaries, particularly when multiple actions in the video exhibit similar visual patterns.

Video:



Question:

The person uses multiple similar objects to play an occlusion game. Where is the hidden object at the end of the game from the person's point of view?

- A. Under the third object from the left.
- B. Under the first object from the left.
- C. Under the second object from the left.

SDRL:

<summary> The person stands behind a small round table with three bowls and a can. They move the bowls around, seemingly playing a game of hiding an object under one of the bowls. After a series of moves, the person lifts the bowls to reveal the can underneath. </summary>

<think> The person's actions suggest they are playing a game of hiding an object under one of the bowls. The sequence of lifting and moving the bowls indicates that the hidden object is likely under one of the bowls that was moved last. Since the person lifts the bowls to reveal the can, it implies that the can was the hidden object. The can was initially placed on the table and then moved under one of the bowls, which the person then lifted to reveal it. </think>

<answer> A </answer> ❌

Figure 9. A failure case from MVBench dataset. When actions become highly fine-grained, SDRL often generates coarse, high-level summaries instead of enumerating precise motion steps.