

## A. Training Details

We build our training pipeline on VERL and vLLM 0.10. We recommend using these exact versions for reproduction, as different vLLM versions affect training and inference performance and may cause memory leaks. Note that we extend the original VERL framework to support mixed image and video training, as the vanilla VERL implementation does not natively handle multimodal inputs with heterogeneous visual formats.

**Model and Algorithm.** We use Qwen2.5-VL-7B-Instruct as the base model and employ GRPO (Group Relative Policy Optimization) with advantage normalization. We disable the KL penalty during training and set the PPO clip ratio to 0.2 and the value clip ratio to 3.0.

**Data Configuration.** We set the maximum prompt length to 8,192 tokens and the maximum response length to 4,096 tokens. We use a training batch size of 128 and a validation batch size of 64.

**Training Hyperparameters.** We train for 2 epochs with a constant learning rate of  $1 \times 10^{-6}$  without warmup and weight decay of 0.01. The PPO mini-batch size is 64 and the micro-batch size is 2 per GPU. We apply gradient clipping with a maximum norm of 1.0 and perform 1 PPO epoch per iteration.

**Sampling Strategy.** During rollout, we generate 8 responses per prompt using vLLM with temperature 1.0 and top- $p$  sampling of 1.0. We disable greedy top- $k$  sampling ( $k = -1$ ).

**Distributed Training.** We apply FSDP (Fully Sharded Data Parallel) with a sharding size of 2 for the actor model. The reference model uses FSDP with parameter offloading enabled. Model checkpoints are saved every 20 iterations and include the model weights, optimizer states, and HuggingFace-compatible model files.

## B. Training Efficiency Comparison

We compare the training efficiency of VideoVAST against existing video reasoning methods in terms of training data size and computational cost. Table 1 presents a comprehensive comparison with Video-R1 [1] and VideoRFT [2], which represent state-of-the-art video RL approaches. All compute is reported in A100-equivalent GPU hours.

Table 1. Training efficiency comparison across methods.

Method	SFT	RL	SFT Data (K)	RL Data (K)	Total Data (K)	GPU Hours
Video-R1 [1]	✓	✓	165	260	425	440
VideoRFT [2]	✓	✓	102	310	412	–
VideoVAST	✗	✓	–	15	15	123

Table 1 shows that VideoVAST achieves significant ef-

iciency improvements over existing methods. Compared to Video-R1, VideoVAST reduces the training dataset from 425K samples to 15K samples (96.5% reduction) and GPU hours from approximately 440 to 123 (72.0% reduction). Both Video-R1 and VideoRFT require separate supervised fine-tuning and reinforcement learning stages with large-scale data. In contrast, VideoVAST trains directly with reinforcement learning on ability-stratified data and eliminates the supervised fine-tuning initialization stage.

VideoRFT uses 412K training samples across both stages. VideoVAST demonstrates that ability-driven data curation enables efficient training with only 15K samples (96.4% reduction). These results validate our hypothesis that ability-driven data organization outperforms massive scaling with random sampling. VideoVAST achieves high data efficiency while maintaining competitive computational performance.

## C. VAST Data Distribution

VAST contains 14,274 video-question pairs spanning three cognitive layers. Table 2 presents the detailed distribution of samples across ten capability categories.

Table 2. Detailed distribution of VAST across capability categories.

Layer	Capability	Samples	Percentage
Perception	Spatial Entity Perception	2,000	14.0%
	Spatio-Temporal Dynamics	3,000	21.0%
Reasoning	Causal Reasoning	1,250	8.7%
	Intentional Reasoning	1,250	8.7%
	Procedural Reasoning	1,250	8.7%
	Relational Reasoning	524	3.7%
Cognition	Rule Discovery	2,000	14.0%
	Symbol Comprehension	1,000	7%
	Theme Inference	1,000	7%
	Expression Analysis	1,000	7%
<b>Total</b>		<b>14,274</b>	<b>100%</b>
Perception (Layer 1)		5000	35.0%
Reasoning (Layer 2)		4274	30.0%
Cognition (Layer 3)		5000	35.0%

As shown in Table 2, VAST contains 14,274 samples distributed across three cognitive layers. The dataset comprises 35.0% perception-level samples, 30.0% reasoning samples, and 35.0% cognition samples. Within the perception layer, spatial entity perception (14.0%) and spatio-temporal dynamics (21.0%) constitute the primary capabilities. The reasoning layer includes causal reasoning (8.7%), intentional reasoning (8.7%), procedural reasoning (8.7%), and relational reasoning (3.7%). The cognition layer en-

compasses rule discovery (14.0%), symbol comprehension (7%), theme inference (7%), and expression analysis (7%). Compared to existing datasets like Video-R1-260K, which contains approximately 82% perception samples, 15% reasoning samples, and only 3% cognition samples, VAST provides substantially more balanced coverage of higher-order reasoning and abstract cognition capabilities. It is worth noting that, because the Video-R1 data for Relational Reasoning is limited to only 524 instances, we sampled all available examples, even though this number is not consistent with the sample sizes of the other Layer 2 categories.

## D. Additional Ablation Studies

### D.1. KL Penalty vs Consistency Reward

We compare the performance of using KL penalty with different  $\beta$  values versus our proposed consistency reward mechanism. The KL penalty is a standard regularization technique in GRPO that constrains the policy deviation from the reference model. We evaluate two KL penalty configurations ( $\beta = 0.02$  and  $\beta = 0.04$ , which are utilized in Video-R1) and compare them with our consistency reward approach ( $\lambda = 0.7$ ). Table 3 presents the results across seven video understanding benchmarks.

Table 3. Comparison of KL penalty and consistency reward.

Method	VSI-Bench	Video MMMU	MMVU (mc)	MV Bench	Temp Compass	VideoMME (wo sub)	Video Holmes	Overall
w/ KL penalty ( $\beta = 0.02$ )	<u>33.4</u>	48.2	<u>65.1</u>	65.8	73.4	<u>57.6</u>	37.1	54.4
w/ KL penalty ( $\beta = 0.04$ )	31.4	<u>49.8</u>	65.0	<b>66.7</b>	<b>74.8</b>	<b>57.7</b>	<u>37.8</u>	<u>54.7</u>
w/ CR (ours)	<b>33.5</b>	<b>50.6</b>	<b>66.1</b>	<u>66.3</u>	<u>74.2</u>	<b>57.7</b>	<b>42.5</b>	<b>55.8</b>

Table 3 demonstrates that our consistency reward mechanism achieves better performance compared to the traditional KL penalty. Our approach obtains an overall score of 55.8, outperforming the KL penalty baseline (54.7) by 1.1. Consistency reward achieves the best performance on 5 out of 7 benchmarks, with notable improvements on VSI-Bench (33.5 vs. 31.4), VideoMMMU (50.6 vs. 49.8), MMVU (66.1 vs. 65.0), and particularly Video-Holmes (42.5 vs. 37.8). The substantial gain on Video-Holmes (4.7) is especially significant, as this benchmark emphasizes complex temporal reasoning and causal understanding. While KL penalty slightly outperforms consistency reward on MV Bench (66.7 vs. 66.3) and TempCompass (74.8 vs. 74.2), these differences are marginal. These results validate our design choice to replace the KL penalty with consistency rewards, demonstrating that explicit reasoning-answer coherence regularization is more effective for video understanding than generic policy constraints.

### D.2. Consistency Reward Weight

In our consistency-regularized reward design (Eq. 3), the consistency reward is defined as  $R_{cr} = \lambda \times s_{consistency}$ , where

$s_{consistency} \in \{0, 1\}$  is a binary signal computed by a consistency model that evaluates alignment between the reasoning process (within `<think>` tags) and the final answer. We investigate the impact of different weight values for  $\lambda$  while keeping  $R_{format} = 0.5$  and  $R_{acc} = 1.0$  fixed. Table 4 presents the results across seven video understanding benchmarks.

Table 4. Ablation study on consistency reward weight  $\lambda$ . We use **bold** to indicate the best performance and underline for the second best.

$\lambda$	VSI-Bench	Video MMMU	MMVU (mc)	MV Bench	Temp Compass	VideoMME (wo sub)	Video Holmes	Overall
0.5	32.1	50.4	<u>64.6</u>	65.8	<b>74.1</b>	56.8	41.2	55.0
0.7	<b>33.5</b>	50.6	<b>66.1</b>	<b>66.3</b>	<u>74.2</u>	<b>57.7</b>	<b>42.5</b>	<b>55.8</b>
1.0	<u>32.5</u>	<b>52.7</b>	64.3	<u>66.7</u>	72.9	<u>57.2</u>	<u>41.7</u>	<u>55.4</u>

Table 4 shows that  $\lambda = 0.7$  achieves the best overall performance (55.8), outperforming both lower and higher weight values. When  $\lambda = 0.5$ , performance drops to 55.0, with notable degradation on VSI-Bench (32.1 vs. 33.5), MMVU (64.6 vs. 66.1), and Video-Holmes (41.2 vs. 42.5), indicating that insufficient consistency regularization fails to guide the model toward coherent reasoning patterns. Conversely, increasing  $\lambda$  to 1.0 also reduces overall performance to 55.4, despite achieving the highest VideoMMMU score (52.7). The substantial drop on TempCompass (72.9 vs. 74.2) suggests that excessive consistency weight over-constrains the policy and limits exploration of diverse reasoning paths. This aligns with our design in Sec. 4.2, where we set  $\beta = 0$  to remove the KL penalty and encourage exploration. The optimal weight of 0.7 achieves the best performance on 5 out of 7 benchmarks, balancing reasoning-answer coherence with sufficient exploration capacity.

### D.3. VideoVAST Resolution

We investigate the impact of resolution settings in VideoVAST. Table 5 compares three resolution configurations (128, 256, and 512) across six video understanding benchmarks.

Table 5. Ablation study on VideoVAST resolution. We use **bold** to indicate the best performance and underline for the second best.

Resolution	VSI-Bench	Video MMMU	MMVU (mc)	MV Bench	Temp Compass	VideoMME (wo sub)	Overall
128	33.8	47.7	65.0	64.0	72.0	55.2	56.3
256	<u>33.5</u>	<u>50.6</u>	<u>66.1</u>	<u>66.3</u>	<u>74.2</u>	<b>57.7</b>	<u>58.1</u>
512	<b>35.1</b>	<b>50.4</b>	<b>66.6</b>	<b>66.5</b>	<b>74.7</b>	<u>57.5</u>	<b>58.5</b>

Table 5 shows that higher resolution consistently improves performance. Resolution 512 achieves the best overall score (58.5), outperforming resolution 256 (58.1) and resolution 128 (56.3). This configuration delivers the best results on 5 out of 6 benchmarks, with substantial gains

on VSI-Bench (35.1 vs. 33.5 vs. 33.8) and TempCompass (74.7 vs. 74.2 vs. 72.0). However, resolution 256 achieves the highest score on VideoMME (57.7), slightly outperforming resolution 512 (57.5). This indicates that while higher resolution benefits fine-grained visual understanding and temporal reasoning tasks, a trade-off exists between computational efficiency and performance gains on certain benchmarks.

#### D.4. Curriculum Learning Strategy

We investigate the impact of different curriculum learning strategies on model performance. We compare four training strategies: (1) Sequential trains each level separately in sequence (L1→L2→L3), (2) Two-Stage trains L1 first, then jointly trains L2 and L3, (3) Smooth gradually transitions from L1 to L3 with weighted sampling, and (4) Uniform jointly trains all levels with uniform sampling (L1+L2+L3). Table 6 presents the results across seven video understanding benchmarks.

Table 6. **Ablation study on curriculum learning strategy.** We use **bold** to indicate the best performance and underline for the second best.

Strategy	VSI-Bench	VideoMMU	MMVU (mc)	MV Bench	Temp Compass	VideoMME (wo sub)	Video Holmes	Overall
Sequential	32.2	50.3	64.8	64.9	<u>74.4</u>	55.6	35.8	54.0
Two-Stage	<u>35.5</u>	<b>50.6</b>	<u>65.8</u>	66.0	74.3	<u>57.0</u>	<u>38.6</u>	55.4
Smooth	<b>35.8</b>	49.0	65.4	<u>66.1</u>	<b>74.5</b>	<u>57.0</u>	41.2	<u>55.6</u>
Uniform	33.5	<b>50.6</b>	<b>66.1</b>	<b>66.3</b>	74.2	<b>57.7</b>	<b>42.5</b>	<b>55.8</b>

Table 6 presents the performance of four curriculum learning strategies. Uniform achieves the highest overall score of 55.8, excelling on Video-Holmes (42.5), VideoMME (57.7), MMVU (66.1), and MVBench (66.3). This demonstrates that jointly training all difficulty levels with uniform sampling provides balanced capability development across different reasoning complexities. Smooth ranks second with an overall score of 55.6, achieving the best performance on VSI-Bench (35.8) and TempCompass (74.5). This strategy gradually transitions from L1 to L3 through weighted sampling, enabling the model to progressively adapt to more complex reasoning tasks. Two-Stage reaches 55.4 overall, ranking second on VSI-Bench (35.5) and Video-Holmes (38.6), indicating that initial learning on simpler tasks followed by combined training on more complex levels offers a reasonable compromise. The Sequential strategy yields the lowest overall score (54.0), suggesting that strict progressive training through difficulty levels provides limited benefit.

## E. Prompt Template

### E.1. System Prompt

We use the following system prompt to guide the model to generate reasoning within `<think>` tags and answers

within `<answer>` tags:

```
A conversation between User and Assistant. The user asks a question,
and the Assistant solves it. The assistant first thinks 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>.
```

### E.2. Question Type Templates

We define task-specific templates for different question types:

```
Multiple Choice:
Answer the following question: {question}
Please reason step by step within <think> </think> tags, and provide
only the single option letter (\eg, A, B, C, D, etc.) within the
<answer> </answer> tags.

Numerical:
Answer the following question: {question}
Please reason step by step within <think> </think> tags, and provide
the numerical value (\eg, 42 or 3.14) within the <answer> </answer>
tags.

OCR:
Answer the following question: {question}
Please reason step by step within <think> </think> tags, and
transcribe text from the image/video clearly and provide your text
answer within the <answer> </answer> tags.

Free-form:
Answer the following question: {question}
Please reason step by step within <think> </think> tags, and provide
your text answer within the <answer> </answer> tags.

Regression:
Answer the following question: {question}
Please reason step by step within <think> </think> tags, and provide
the numerical value (\eg, 42 or 3.14) within the <answer> </answer>
tags.
```

### E.3. Data Annotation Prompt

We use the following prompt for annotating capability categories during dataset construction. The annotator classifies each question into one of 10 capability categories across three cognitive levels (Perception, Reasoning, and Cognition):


```
You are an annotator for capability classification in {task_description}.
Based solely on "{content_description}" + question/options + (optional)
thought process + correct answer, determine the **minimum capability
required** to answer the question, and select **exactly one** category
from the 10 below. Output using the specified JSON template.

[INPUT]
- {question_label}: {question}
- Options: {options_text}
- Thought process (may be empty): {process}
- Correct answer: {answer}
- Explanation: {explanation}

[KEY PRINCIPLES]
1) Minimal sufficiency: Choose the **lowest-level capability** needed
to answer correctly (if perception suffices, do not escalate to
reasoning/cognition).
2) Consider only information required to solve the question; do not
classify content in the {content_reference} that is irrelevant to
this question.
3) Output **strict JSON** with no extra characters/explanations/
Markdown.
4) "reasoning" should include only **conclusive justification**
(<=3 sentences); **do not** reveal step-by-step reasoning/chain-of-
thought.
5) "evidence" is recommended to include **{evidence_example}**; if
timestamps are unknown, still provide observable evidence.
6) "confidence" is in [0,1], keep two decimal places.

[CAPABILITY CATEGORIES (10)]

Perception
1. spatial_entity_perception
Analyze a scene at a **single moment**: object recognition,
segmentation, localization, pose, scene layout; answers "what/where/
what does it look like."
```



**Question:** What is the direct cause of the wound on the woman's neck?

**Options:**

A. Steak knife accidentally cut.

B. Car accident impact caused.

C. self-harm behavior.


D. The man killed with a knife. ✓

E. Supernatural power cutting

F. Red wine bottle fragments pierce in.

**VideoVAST :** <think>The video sequence shows a man holding a knife, which is a direct indication of the cause of the wound on the woman's neck. The presence of the knife and the man's actions suggest that the wound is a result of him using the knife.</think><answer>D</answer>

**Video-R1 :** <think> Hmm, it seems like the most plausible option is A. Steak knife accidentally cut. The man is holding a knife, and the woman is lying on a table, which suggests that the wound might have been caused by an accidental cut from a steak knife.</think><answer>A</answer>



**Question:** What does the focus on the white, orange and blue coat on the ground at 2:04 imply?

**Options:**

A. The owner has been killed. ✓

B. Clothes are discarded.

C. Shooting props bloopers.


D. The weather suddenly turns cold.

E. Clothes were stolen.

F. The owner changes the style.

**VideoVAST :** <think>The focus on the white, orange and blue coat on the ground at 2:04 implies a significant event, given the context of the video which seems to be a dramatic or intense scene. The presence of the coat on the ground, along with the other elements of the scene, suggests a possible murder or death, as this is a common prop in horror or thriller genres to indicate a violent event.</think><answer>A</answer>

**Video-R1 :** Option C suggests that it's a shooting props blooper. This is an interesting option. In film production, sometimes props are left on the ground by mistake, and the focus on the coat could be highlighting this blooper. This seems like a plausible explanation.<answer>C</answer>



**Question:** What was the intention of the man in black aiming his gun at the man in red at 0:20?

**Options:**

A. Self-defense counterattack.

B. Threatening to extort money.

C. Misidentification.


D. Prevent alarm.

E. Vent anger.

F. carry out a cover-up. ✓

**VideoVAST :** <think>The man in black is aiming his gun at the man in red, which suggests a threatening action. Given the context of the scene, it is likely that the man in black is threatening to extort money from the man in red.</think><answer>F</answer>

**Video-R1 :** <think> The most plausible intention of the man in black is to threaten to extort money from the man in red. This fits the context of the scene where the man in black is holding a gun and the man in red is on the ground.</think><answer>B</answer>



**Question:** Which item is closest to the core theme of this film?

**Options:**

A. The wisdom of the weak rising up.

B. The awareness of natural ecological protection.

C. The preciousness of innocent childhood.

D. The importance of exercise.

E. Those who harm others will ultimately bring misfortune upon themselves. ✓

F. The warning of out-of-control technology.

**VideoVAST :** <think>The film appears to be a fable or a story that involves a character who is initially in a position of power or control but ends up being punished for their actions. This is a common theme in fables where the moral is often about the consequences of one's actions. The other options are either too broad or specific to the context of the film, but the theme of punishment for harming others is a central element in the narrative structure of the film.</think><answer>E</answer>

**Video-R1 :** <think> The film shows a young boy running through a park, interacting with nature, and eventually falling asleep on the grass. The boy's actions and the setting suggest a focus on childhood and innocence.</think><answer>C</answer>

Figure 1. **Qualitative comparison between VideoVAST and Video-R1.** VideoVAST exhibits more temporally grounded and causally coherent reasoning traces than Video-R1, suggesting reduced reliance on shortcut cues.

<div> <div>2. spatio_temporal_dynamics</div> <div>Analyze <b>changes over a time interval</b>: identity and trajectory tracking, pose/attribute changes (e.g., a light gets brighter), atomic/complex action recognition, and event start/end boundaries.</div> <div>Reasoning</div> <div>3. causal_reasoning</div> <div>Infer physical/logical chains between events; requires cause-effect or necessary/enabling conditions (may include counterfactuals).</div> <div>4. intentional_reasoning</div> <div>Infer an agent's goals, beliefs, and plans (mental states) from behavior, e.g., "reaching for a cup is to drink water."</div> <div>5. procedural_reasoning</div> <div>Multi-step task order and conditional dependencies; detect missing/incorrect steps or parallel/branching.</div> <div>6. relational_reasoning</div> <div>From multi-agent interactions, infer roles, status, power, and other social relations (cooperation/competition/norms, etc.).</div> <div>Cognition</div> <div>7. rule_discovery</div> <div>Identify underlying rules of the scene (physical or social) and detect violations/anomalies.</div> <div>8. symbol_comprehension</div> <div>Interpret figurative/symbolic meanings and cultural references of visual elements (beyond literal semantics).</div> <div>9. theme_inference</div> <div>Extract the central idea/moral/message as a global summary of the storyline.</div> <div>10. expression_analysis</div> <div>Analyze audiovisual language (color, composition, camera movement, etc.) and how it conveys emotion, sets tone, or forms style.</div> <div>[ADJUDICATION RULES (use these for common confusions)]</div> <div>- spatial_entity_perception vs spatio_temporal_dynamics:</div> <div>If the answer depends only on <b>single-frame/instantaneous</b> category, location, appearance, pose, or layout → the former; if it</div> </div>	<div> <div>requires <b>cross-frame consistency/trajectory/change/action or event boundaries</b> → the latter.</div> <div>- spatio_temporal_dynamics vs procedural_reasoning:</div> <div>If it's merely recognizing <b>whether an action/event occurs and when it starts/ends</b> → spatio-temporal; if it requires understanding <b>multi-step order and prerequisites</b> → procedural.</div> <div>- causal_reasoning vs mere temporal order:</div> <div>"Before/after" alone != causation; assign <b>causal</b> only when <b>necessary/sufficient/enabling conditions or intervention cues</b> are needed.</div> <div>- intentional_reasoning vs causal_reasoning:</div> <div>If it requires explaining <b>agent goals/intentions/plans</b> → intentional; if it's about <b>objective physical/logical causation</b> (independent of mental states) → causal.</div> <div>- intentional_reasoning vs relational_reasoning:</div> <div>If the judgment depends on <b>social roles/norms/power structures</b> → relational; if only individual intention is needed → intentional.</div> <div>- rule_discovery vs procedural_reasoning:</div> <div><b>Rules</b> are general/normative (discover regularities and violations); <b>procedure</b> is the step dependencies of a specific task instance.</div> <div>- theme_inference vs symbol_comprehension vs expression_analysis:</div> <div><b>Theme</b> = the work's central idea; <b>symbol</b> = the symbolic meaning of objects/colors/actions; <b>expression</b> = audiovisual techniques and their effect on mood/style.</div> <div>[OUTPUT FORMAT (strict JSON; select only one best-fitting category)]</div> <div>{</div> <div>"summary_label": "{summary_description}",</div> <div>"primary_category": "spatial_entity_perception spatio_temporal_dynamics causal_reasoning intentional_reasoning procedural_reasoning relational_reasoning rule_discovery symbol_comprehension theme_inference expression_analysis",</div> <div>"category_name": "Name (corresponding to primary_category)",</div> <div>"level": "Perception Reasoning Cognition",</div> <div>"reasoning": "Conclusive justification in no more than 3 sentences (no step-by-step reasoning).",</div> <div>}</div> </div>
--	--



(a)

**Question:** In the illustrated circuit, the consumption power of load  $R_L$  is 72W. If the transformer voltage ratio  $k=5$ , the effective value of the transformer is  $U_1$  (?) ?

**Options:**

A. equal to 120V ✓  
 B. equals to 4.8V  
 C. and equals to 0.6V  
 D. cannot be calculated  
 E. is 24V  
 F. equals to 60V  
 J. and is 12V  
 H. equals to 240V  
 I. is 48V  
 K. equals to 600V

**VideoVAST:** <think>The question asks for the effective value of the transformer voltage  $U_1$ , given that the consumption power of the load  $R_L$  is 72W and the transformer voltage ratio  $k=5$ ... The question does not provide the current  $I$ , so we cannot calculate the effective value of the transformer voltage  $U_1$ . The correct answer is the one that matches the given options and is consistent with the given information.</think><output>J</output>

(b)

**Question:** Which arrow in points to a large ribosomal unit?

**Options:**

A. A C. C E. E G. G I. I ✓  
 B. B D. D F. F H. H

**VideoVAST:** <think>The question asks which arrow points to a large ribosomal unit. The image shows a diagram of a ribosome with two main parts: the large subunit and the small subunit. The large subunit is labeled with the number 50S. The arrows in the diagram point to different parts of the ribosome. The arrow labeled C points to the large subunit of the ribosome, which is the 50S unit. The other arrows point to different parts of the ribosome, but C is the one that specifically points to the large ribosomal unit.</think><output>C</output>

(c)

**Question:** What is the action performed by the person in the video?

**Options:**

A. Collecting something behind something.  
 B. Spilling something behind something. ✓  
 C. Not sure.

**VideoVAST:** <think>The person in the video is moving a black suitcase from the floor to the right side of the couch. The suitcase is being pushed or dragged, and it is clear that the person is collecting the suitcase behind the couch. The action is a deliberate movement of the suitcase, which is a form of collecting something behind something.</think><answer>A</answer>

(d)

**Question:** What is the action performed by the person in the video?

**Options:**

A. Catching something onto a surface.  
 B. Not sure.  
 C. Throwing something onto a surface. ✓

**VideoVAST:** <think>The video shows a pair of hands holding a piece of paper and folding it. The action involves the hands bringing the edges of the paper together, which is a common action when catching something onto a surface. The hands are positioned in a way that suggests the paper is being caught or held in place, rather than being thrown. The action is deliberate and controlled, indicating that the paper is being caught onto a surface, likely a table or a similar flat surface.</think><answer>A</answer>

Figure 2. **Additional qualitative analysis of failure cases.** Representative examples of typical failure patterns of our method in challenging scenarios.

```
"evidence": [{"Evidence 1 (preferably with {evidence_example})",
"Evidence 2"},
"question_analysis": "Explain why this case belongs to the chosen
category based on the question and answer; state the key information
required."
}

Begin your analysis now and output only the JSON above. If
{options_text} is empty, ignore statements related to options, but
keep all other requirements unchanged.
```

## E.4. Consistency Reward

```
You are a strict, impartial AI grader. Your sole task is to
evaluate the internal consistency between a model's reasoning
process ('<think>') and its final chosen answer. You must not judge
whether the answer is factually correct.

## Task
Compare the final conclusion of the '<think>' block with the
provided "Chosen Answer Text". Determine if the reasoning logically
and unambiguously supports the chosen answer.

## Inputs
1. Problem:
{raw_prompt}
2. Model's Reasoning ('<think>')**:
{think}
3. Model's Chosen Answer Key: '{answer_key}'
4. Chosen Answer Text: '{answer_text}'

## Scoring Criteria
You must assign a score based on the following strict rules:

### Score 1.00 if:
- The '<think>' block's final conclusion clearly and explicitly
states or strongly implies the "Chosen Answer Text".
- The reasoning process, as a whole, is logically aligned with the
final choice. Paraphrasing is acceptable as long as the core
meaning is identical.
- Example of a clear conclusion in '<think>': "Based on this
analysis, option C is the correct one.", "Therefore, the answer is
C.", "This leads me to conclude C."

### Score 0.0 if ANY of the following conditions are met:
- The final conclusion in '<think>' supports a different option
(e.g., reasoning points to B, but the answer is C).
- The reasoning contradicts the chosen answer.
```

```
- The reasoning is vague, inconclusive, irrelevant, or offers
multiple conflicting conclusions without a final decision.
- The '<think>' block is empty or missing a coherent thought
process.
- The "Chosen Answer Text" could not be found or is empty (this is
an input issue, but should be scored as inconsistent).

## Output Format
- You MUST output ONLY the numerical score.
- Your entire response must be either '1' or '0'.
- DO NOT include any explanation, justification, or extra text.

Response:
```

## F. Limitations and Future Works

While our work demonstrates effective data-efficient training for video reasoning, several limitations remain. First, our current approach relies on uniform frame sampling from videos. This strategy may miss critical temporal events or actions that occur between sampled frames. Adaptive frame sampling methods that identify and prioritize informative frames could further improve reasoning quality while maintaining efficiency. Second, our annotation process for capability categorization requires manual effort. In future work, we will develop automated annotation methods, for example, by leveraging strong foundation models to classify capabilities and generate reasoning traces, which will enable larger-scale dataset construction while maintaining quality.

## G. Failure Cases

Figure 2 presents representative failure cases that fall into two main categories: genuine reasoning errors and issues

arising from annotation noise or semantic ambiguity in the ground-truth labels. Cases (a) and (b) represent genuine model failures: (a) incorrectly concludes that the transformer circuit problem is unsolvable despite being answerable through multi-step reasoning; (b) exhibits a spatial grounding error by misidentifying the ribosome diagram arrow (option C vs. correct label I). Cases (c) and (d) reflect label ambiguity: (c) predicts “collecting the suitcase” vs. ground truth “spilling”; (d) predicts “catching onto a surface” vs. ground truth “throwing”. These examples reveal both genuine reasoning failures and annotation quality issues that may cause accuracy metrics to underestimate model performance.

## References

- [1] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-rl: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. [2](#)
- [2] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorf: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. [2](#)