

# Alternating Perception-Reasoning for Hallucination-Resistant Video Understanding

Bowei Pu Chuanbin Liu\* Yifan Ge Peicheng Zhou Yiwei Sun  
 Zhiying Lu Jiankang Wang Hongtao Xie†  
 University of Science and Technology of China, Hefei, China  
 {pubowei}@mail.ustc.edu.cn, {liucb92, htxie}@ustc.edu.cn

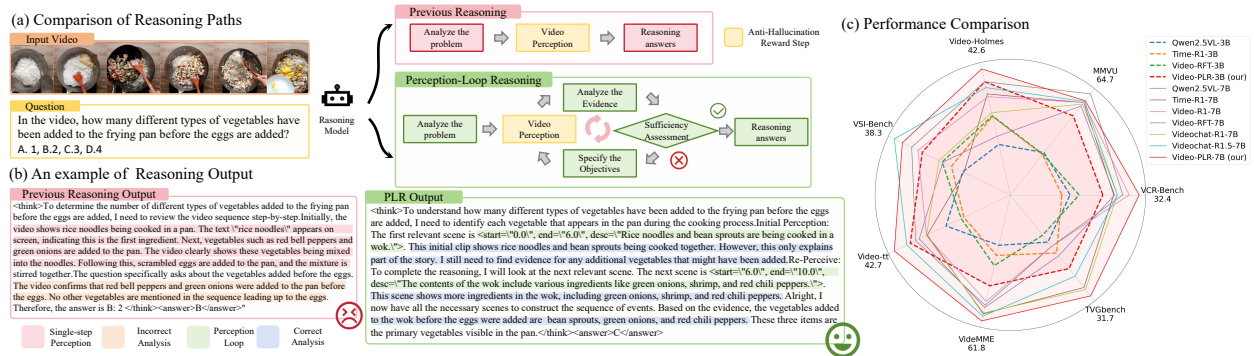


Figure 1. **Overview of VideoPLR.** (a) Compared with reasoning paradigms. VideoPLR performs iterative perception and reasoning, and each perception execution targets the perception results of the previous reasoning. During training, it is allowed to individually judge whether hallucinations occur in each perception result and provide rewards. (b) Examples of VideoPLR. The PLR paradigm perceives partial video clips multiple times, achieving repeated perception and providing correct answers. (c) Benchmark performance: Comparative results on 7 benchmarks highlight the video reasoning capabilities of our model, at both 3B and 7B parameter scales.

## Abstract

Sufficient visual perception is the foundation of video reasoning. Nevertheless, existing Video Reasoning LLMs suffer from perception shortcuts, relying on a flawed single-step perception paradigm. This paradigm describes the video and then conducts reasoning, which runs the risk of insufficient evidence and emergent hallucinations. To address these issues, we introduce a new framework that integrates a loop-based paradigm with an anti-hallucination reward. First, to address the insufficient evidence, we introduce the Perception Loop Reasoning (PLR) paradigm. Instead of describing the video at once, each loop requires the model to describe a video segment with precise timestamps, analyze this segment, and decide the next action. Second, for the risk of hallucinations, the Factual-Aware Evaluator (FAE) evaluates each perception result as a reliable anti-hallucination reward. This reward encourages the model to provide sufficient and precise video evidence. Our FAE, which performs comparably to GPT-4o, is tuned on our AnetHallu-117K, a large-scale hallucination judgment preference dataset. Extensive experiments show that our Video-PLR achieves the state-of-the-art in

both 3B and 7B parameter scales and has the best data efficiency. Our code, datasets, and models is released on: <https://github.com/BoweiPu/VideoPLR>.

## 1. Introduction

Video Large Language Models (VideoLLMs) [21, 33, 45] demonstrate remarkable potential in traditional video understanding. Building on this success, researchers are increasingly turning their attention to complex video tasks. Inspired by the *Think-then-Answer* paradigm [17], many VideoLLMs [12, 32, 43] tackle complex problems by generating intermediate reasoning steps. This reasoning is based on *video perception*: the accurate description of visual information in the video. Thus, insufficient perception or hallucinations often lead to flawed or erroneous reasoning.

The previous paradigm is to describe the video first and then reason. This paradigm has fundamental flaws. As shown in Fig. 1(b), the video description in the *Previous Reasoning* overlooks the video segment of adding bean sprouts, losing the key evidence. Also, this reasoning is inaccurate and exhibits hallucinations. This example shows that the single-step perception faces two challenges: (1)

**Partial and Incomplete Perception:** Single perception often analyzes a part of the video, lacking a mechanism to assess the sufficiency of the evidence. This leads to the loss of key evidence. (2) **Risk of Hallucinatory Perception:** The entire video contains rich and dynamic visual information, making it difficult to assess the accuracy of the overall description. This leads to a loss of reliable hallucination supervision.

Based on the analysis of single-step perception, we argue that a better paradigm should enable the model to repeatedly perceive different video segments and provides reliable hallucination supervision during training. The number of perceptions should be decided automatically by the model itself. Thus, we introduce **Perception Loop Reasoning (PLR)** paradigm, which reformulates this single-step perception into a multi-step paradigm, as shown in Fig. 1(a). The core idea of PLR is to explicitly decouple *Perception* and *Reasoning* into repeating steps. Different from traditional video describing, PLR requires the model to provide timestamps and corresponding descriptions. After each perception, the model will analyze this segment and decide whether to output an answer or continue to perceive. In this way, our PLR paradigm ensures that the completeness of the reasoning is grounded in accurate and sufficient perceptual evidence gathered iteratively. Fig. 1(b) demonstrates multi-step and precise perception.

Realizing this multi-step perception is challenging, as current models lack the ability to re-perceive, and no dataset exists to support it. Following the PLR paradigm, we construct **VideoPLR-14K**, a novel dataset designed to bootstrap the model to generate responses in the PLR format, encompassing both sequential causal reasoning and non-sequential segment ordering. Furthermore, anti-hallucination is the key to precise perception, which is the foundation of reasoning. However, previous models, including Open-source VideoLLMs [22, 34] and vision-language models [42], cannot provide reliable factual evaluation. To fill this gap, we introduce a **Factual-Aware Evaluator (FAE)**, which is meticulously trained on our newly constructed **AnetHallu-117K** dataset. The ANetHallu-117K is constructed from the human-annotated ANet [10] and contains 5 types of hallucination and hybrid CoT binary classification preference data, which can reduce discriminative bias. FAE is comparable to GPT-4o in evaluating text-video consistency to serve as a reliable anti-hallucination reward for RL training. The segmented approach of PLR is also more reliable for hallucination evaluation. Specifically, their constrained visual context simplifies grounding the descriptions, making anti-hallucination more readily identifiable.

These two efforts jointly constitute our final model, **Video-PLR** via reinforcement learning [17], achieving high data efficiency. For the RL stage, we utilize only **6K** of

mixed data, including five types of tasks. Our models (3B and 7B) achieve SOTA results on multiple benchmarks, as illustrated in Fig. 1(c). Compared with previous models, Video-PLR integrates accurate timestamps directly into the reasoning text, enabling a unified consideration of temporal grounding. It is worth mentioning that by simply adding 0.5K of Video Temporal Grounding data, we outperform models trained with more than 5 times the amount of data. PLR not only emphasizes the necessity of sufficient evidence but also takes into account time perception, visual content perception, and logical reasoning, aiming to build a more advanced video reasoning paradigm.

The contributions of this paper are as follows:

- We propose **Perception Loop Reasoning (PLR)**, a novel video reasoning paradigm that addresses the perception shortcut. PLR reformulates video reasoning from a single-step process into a multi-step perception and reflection loop, where the model iteratively perceives segments and analyzes them before concluding.
- To enable fine-grained hallucination evaluation within the loop, we introduce a **Factual-Aware Evaluator (FAE)** and its corresponding large-scale preference dataset, **AnetHallu-117K**. The FAE reaches the level of GPT-4o in hallucination benchmarks and generates anti-hallucination rewards in RL training.
- We construct **VideoPLR-14K**, a dataset designed to bootstrap the PLR paradigm. Our resulting model, **Video-PLR**, trained via RL using FAE as an anti-hallucination reward, achieves SOTA performance both on 3B and 7B parameter scales with data efficiency.

## 2. Related Work

### 2.1. Multimodal Chain-Of-Thought

Multimodal Chain-of-Thought (MCoT) simulates human reasoning by decomposing complex problems into intermediate steps, thereby enhancing transparency and performance. LLaVA-CoT [39] is the first to implement a multimodal O1-like [25] reasoning approach. Similarly, Video-of-Thought [11] proposes a comprehensive five-stage framework. VideoEspresso [18] and VideoCoT [35] achieve MCoT by invoking external tools as a multi-stage approach. Chain-of-Frames [16] further integrates this capability into the model’s internal thinking process, enabling it to describe keyframe content. However, single-frame descriptions are static, making it difficult to perform dynamic analysis across frames. We argue that the video modality necessitates CoT because videos contain vast amounts of information, requiring the model to identify crucial details while filtering out noise. The model’s process of progressively identifying key information is inherently dynamic, and our designed perception and analysis loop supports this process.

## 2.2. Reasoning VideoLLMs

VideoR1 [12] first achieves R1-like [17] video reasoning and utilizes GRPO [29] for training. However, its cold-start data is sampled from Qwen2.5VL 72B [1]. This strategy overlooks the fact that even models with larger parameter counts lack video reasoning training. To enhance video perception, VideoR1 and DeepVideo-R1 [26] enhance the model’s perception by modifying the standard GRPO. VideoRFT [32] advances this further by restructuring the CoT into a fixed pipeline: task, perception, reasoning, and conclusion. It utilizes SigLip [31] during the perception stage to mitigate hallucinations. TimeR1 [36] and VideoChatR1 [23] place a greater emphasis on temporal perception and often lack consideration for general-purpose tasks. VideoChatR1.5 [40] achieves tool-use for video trimming based on temporal reasoning. ARCHunYuanVideo [15] can generate dense video descriptions with timestamps, but lacks the ability to reason within these segments, and also does not address the hallucination problem. In contrast, our reasoning paradigm integrates temporal perception and execution, anti-hallucination training, and cyclical evidence gathering, representing a more advanced approach to reasoning training.

## 2.3. Video Hallucination

Longer reasoning chains reduce attention to visual input, thereby leading to hallucinations [24]. Consequently, hallucination control is crucial for video models. VideoRFT [32] utilizes SigLip [31]. Nevertheless, SigLip is not optimally suited for video. Utilizing a VideoLLM as a judge requires the model itself to be low-hallucinating and unbiased. However, the current models cannot achieve this. VidHalluc [22] evaluates hallucinations from three key dimensions, while VideoHalluc [34] focuses on the bias of judging the correctness of descriptions. To address this, we provide a controllable data generation pipeline for hallucination pairs and introduce a video hallucination judgement model. Our primary objective is to supply a reliable anti-hallucination reward.

## 3. Factual-Aware Evaluator

To provide reliable consistency between textual descriptions and visual content, Factual-Aware Evaluator is trained on AnetHallu-117K, which is based on human-annotated ActivityNet (ANet) [10]. In this section, we describe in detail the caption construction pipeline and the preference training pipeline, as shown at the top of Fig. 2. The key to this pipeline is to eliminate the bias of language and strengthen the visual judgment of the model.

## 3.1. Re-annotate Dense Caption

ANet contains video content from various scenarios; however, its original annotations are coarse-grained. It is impossible to obtain hallucinated caption pairs from such annotations. Thus, we first generate high-fidelity dense descriptions as ground-truth,  $C_{gt}$ , for each video segment  $[t_1, t_2]$  utilizing Qwen2.5-VL 32B based on original annotations. The original human annotations guided the model’s video description process, achieving near-realistic dense descriptions.

## 3.2. Low-Bias Judgement

This part includes the pipeline for generating low language bias caption pairs and enhancing perception using CoT.

First, the requirement of low language bias means that it is difficult to distinguish whether it is a hallucination caption through language. We devised a two-stage **Hallucinate-then-Correct** pipeline to maintain the language style and sentence length consistency within a caption pair. A broad range of hallucination types brings better robustness. There are **five** hallucination types: (1) *Attribute Modification*, (2) *Quantity Modification*, (3) *Action Substitution*, (4) *Detail Conflation*, and (5) *Temporal Reordering*. By providing the model with  $C_{gt}$  and the hallucination definition, negative captions of the specified hallucination types can be obtained. For the correction stage, provide the model with both  $C_{neg}$  and  $C_{gt}$  simultaneously. The positive captions,  $C_{pos}$ , generated by the model serve as a direct fix for the specific error present in  $C_{neg}$ . Thus, raw negative and positive caption pairs are collated. Fig. 4 presents the caption pairs that have similar language styles and text lengths.

Because of the nature of LLMs, captions exhibit significant differences in vocabulary distribution. It may cause the vulnerability to be discovered during model training and result in the model exploiting this vulnerability. **Text Bias Elimination** step removes these captions containing the significant bias words, which calculates a bias score for each caption using the biased vocabulary sets  $\mathcal{W}_{pos}$  and  $\mathcal{W}_{neg}$ . This biased vocabulary set comes from the word frequency ratio. Relative frequency ratio  $R_{pos/neg}$  is computed for each word  $w$  in the positive and negative sample sets. Specifically, for words in negative caption sets,  $R_{neg}$  is defined as:

$$R_{neg}(w) = \frac{f_{neg}(w)}{\max(f_{pos}(w), 1)}, \quad (1)$$

where  $f(w)$  is the total frequency of word  $w$  in the corresponding set. The biased vocabulary set is defined as the top  $N$  words with the highest  $R_{pos/neg}(w)$  in each caption set. We assign penalty scores  $P(w)$  with these biased words.

$$P_{neg}(w) = \frac{R_{neg}(w)}{\max_{w' \in \mathcal{W}_{neg}} R_{neg}(w')}. \quad (2)$$

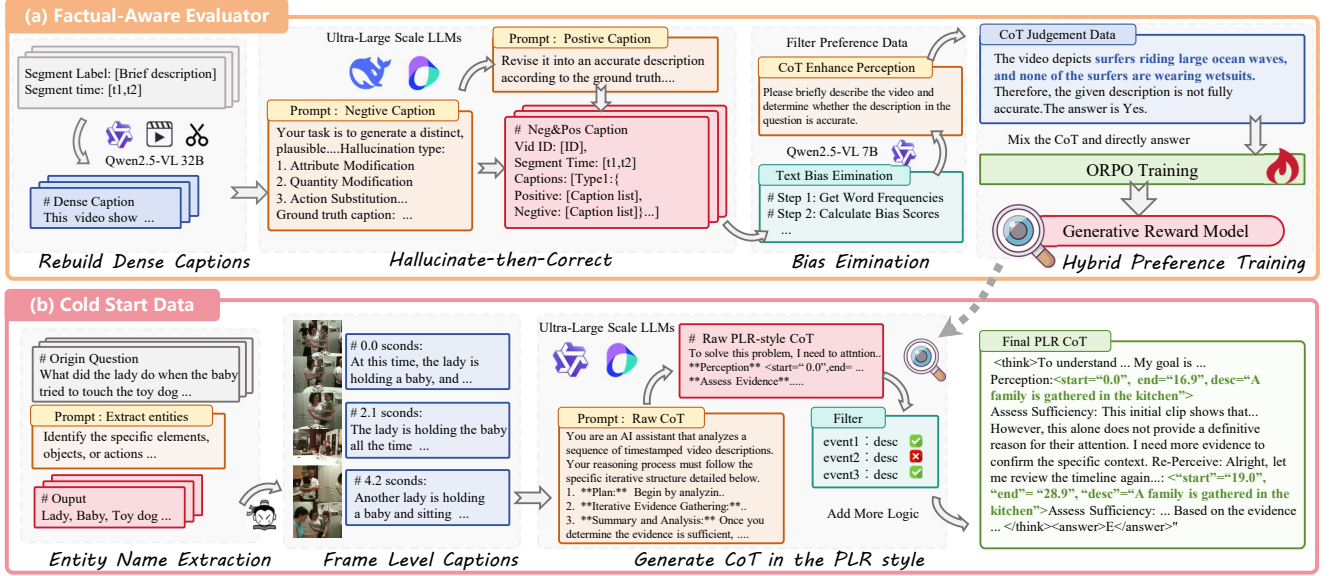


Figure 2. **The pipeline for hallucination preference data and cold start data.** (a) The figure above shows the construction method of hallucination data. By re-annotation, five types of hallucinated captions are generated. Then, Qwen 2.5VL-7B is used to automatically generate CoT data, resulting in the Factual-Aware Evaluator. (b) The figure below shows the pipeline for building cold start data following the PLR paradigm. It mainly demonstrates the method of reannotating NextQA, using carefully designed prompts to enable the model to generate inference data containing perceptual loops.

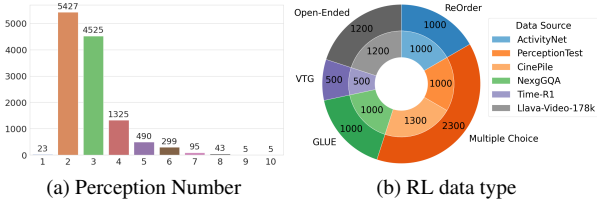


Figure 3. Statistics of the post-training dataset

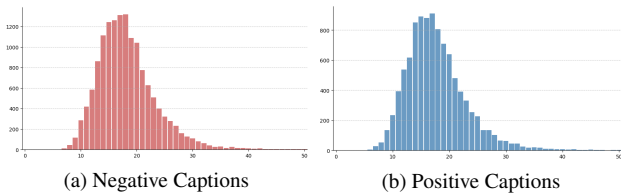


Figure 4. Comparison of caption lengths in ANetHalu-117K

Subsequently, we obtain the total preference score  $S(c_{neg})$  for a given caption  $c_{neg}$  by summing the penalties of all its constituent bias words:

$$S(c_{neg}) = \sum_{w \in W(c_{neg}) \cap W_{neg}} P_{neg}(w), \quad (3)$$

where  $W(c_{neg})$  represents the words in the caption  $c_{neg}$ . We apply the same process to handle positive captions. We finally execute it 15 times, deleting 2% of the data each time, and setting the bias vocabulary set size  $N$  as 30.

Despite word-level frequency bias being eliminated well, the model might still exploit subtle linguistic differences

(e.g., grammatical fluency) between  $C_{pos}$  and  $C_{neg}$  to exploit spurious correlations. Thus, we argue that **CoT Enhance Perception**, which first mandates the model to generate video descriptions. This CoT allows the model to rely on visual details, identify errors, and subsequently arrive at a conclusion.

For a given video clip  $V_{[t_1, t_2]}$  and a caption  $C_{test}$ , we prompt the model to perform a perceptual re-reasoning process: first, generate a brief factual description  $R_{desc}$  of  $V_{[t_1, t_2]}$ , and then, judge whether  $C_{test}$  is correct or incorrect. The final data tuple is formatted as  $(V_{[t_1, t_2]}, C_{test}, R_{desc}, A)$ , where  $A$  is the final binary judgment. We select responses where the model accuracy is neither 0% nor 100% as preference data for training.

### 3.3. Hybrid Preference Training

In the training stage, we construct a mixed training dataset  $\mathcal{D}$ , which contains multiple types of preference pairs  $(x, y_w, y_l)$ , where  $x$  is the input (video-caption pair). The first type is *Answer Accuracy* (Direct Answer vs. Direct Answer), which prioritizes rapid response. The second type is *Reasoning Accuracy* (CoT vs. CoT), compelling the model to engage in perceptual reasoning. This mixed data strategy ensures the model not only judges the correctness of the caption but also learns why, thereby achieving robust and efficient video factual evaluation.

Odds Ratio Preference Optimization (ORPO) [19] is employed for training. The total loss function  $\mathcal{L}_{ORPO}$  for ORPO is defined as a weighted sum of the SFT loss and

the Odds Ratio (OR) preference loss:

$$\mathcal{L}_{ORPO} = \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\mathcal{L}_{SFT} + \lambda \cdot \mathcal{L}_{OR}], \quad (4)$$

where the  $\mathcal{L}_{SFT}$  term,  $\mathcal{L}_{SFT} = -\log P_{\theta}(y_w|x)$ , performs standard supervised fine-tuning exclusively on the winning response  $y_w$ .

The  $\mathcal{L}_{OR}$  term is the preference loss, which penalizes the model’s odds of ranking  $y_l$  higher than  $y_w$ :

$$\mathcal{L}_{OR} = -\log \sigma \left( \log \frac{\text{odds}_{\theta}(y_w|x)}{\text{odds}_{\theta}(y_l|x)} \right). \quad (5)$$

All of these prompts and more details about the dataset can be found in the supplementary materials.

## 4. Perception Loop Reasoning Post-Training

In this section, the Perception Loop Reasoning paradigm is introduced in detail. Then, following this paradigm, we construct SFT data as cold start. In the RL stage, we specifically incorporated an anti-hallucination reward mechanism to encourage the model to generate reasoning chains with accurate facts.

### 4.1. PLR Design

The core concept of PLR involves iteratively describing designated video segments while dynamically evaluating the perceptual process. The PLR-style reasoning is structured into three phases: initial analysis, iterative perception&reasoning, and summarization&conclusion, depicted in the Fig. 1(a). In the first stage, the model should analyze the question independently and determine which visual content may be crucial. The key to PLR lies in the second stage. After each perception, it is necessary to further analyze the relationship between the visual content of the perception and the problem and determine whether it is sufficient to solve the problem. If the problem can be solved, it will directly enter the third stage; otherwise, it will re-execute the perception action, formulate a new perception goal, and output the next perception result. Therefore, this loop is determined autonomously by the model. Each perception follows a specific format, including timestamps and descriptions of the video clip. This allows for independent evaluation of every perception outcome, yielding more granular assessment results compared to full video descriptions. We employ FAE for hallucination evaluation and encourage the model to generate more perception results. Finally, in the third stage, the model synthesizes the perceptual and reasoning content to derive the answer to the question. This special reasoning path requires cold-start data, and next, we will explain how to construct such SFT data from traditional datasets.

### 4.2. Cold-Start Data Construction

Video perception is divided into sequential and disordered ones on the time axis. To accommodate the reasoning for these two sequences, we have covered two types of tasks. For the sequential part, we select NextQA [37], which contains causal and temporal action reasoning. For the disordered part, we set event ordering based on ActivityNet [10] named ReOrdering. The challenge is the lack of timestamp and reasoning annotations within NextQA. Thus, we mainly introduce how to construct the CoT with timestamps of NextQA, as shown at the bottom of Fig. 2(b).

**Entity Name Extraction:** VideoQA suffers from language inconsistency, where synonymous expressions like toy dog and plush toy cause referential ambiguity. To mitigate this, we employ an LLM to extract a normalized entity list  $E$  from the question  $Q$ . This standardizes the terminology and unifies the focus for subsequent steps.

**Frame Level Captioning:** The objective of this step is to acquire detailed visual descriptions for each frame. To maintain descriptive consistency, when captioning frame  $F_t$ , we feed the VLM with the current frame  $F_t$ , the preceding frame  $F_{t-1}$ , the normalized entity list  $E$ , and the previous caption  $C_{t-1}$ . This prompts the VLM to generate dense descriptions focusing on the relevant entities, their behaviors, and the inter-frame changes.

**Raw CoT and Refine Logic:** To obtain the PLR-style CoT from the original description, we divide it into a two-stage optimization. The first stage focuses on generating coherent multi-paragraph descriptions. The second stage focuses on adding logic and self-plan.

For the ReOrdering task, the dense captions in sec 3.2 serve as available evidence. For a video containing multiple events, rearrange the short video descriptions into a ReOrdering task, where the answers are the events arranged in chronological order. For NextQA, we have the basic annotation for each frame, but lack a coherent description of video segments. We feed all descriptions, their timestamps, the question, and the answer into an LLM, tasking it with aggregating these inputs into a structured event list  $V = \{(\text{start}, \text{end}, \text{desc})\}$  within the thinking. Crucially, this step incorporates the FAE to rigorously filter the description of each segment. This process eliminates all inaccurate or fabricated events, ensuring factual accuracy. Although we obtain the Raw PLR-style CoT, there is a lack of logic.

In the second stage, we use carefully crafted prompts to enable the LLM to perform this automated data rewriting. Additionally, at this stage, training samples that do not conform to logic should be removed. Finally, we select 14K high-quality data from the original annotations for cold start, and the perceived numbers are diverse, as shown in Fig. 3a. All prompts and examples of data will be provided in the supplementary materials.

### 4.3. Training with GRPO and Reward Design

The Group Relative Policy Optimization (GRPO) [29] process begins by sampling  $K$  candidate responses  $\{o_i\}_{i=1}^K$  for a given query  $q$  and evaluating their rewards  $\{r_i\}_{i=1}^K$ . The rewards are then normalized to compute the advantage  $A_i$  for each response:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \dots, r_K\})}{\text{std}(\{r_1, \dots, r_K\})}, \quad (6)$$

where `mean` and `std` are calculated over the  $K$  rewards.

The policy  $\pi_\theta$  is then optimized by maximizing the following objective function:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{[q, \{o_i\}]} \frac{1}{K} \sum_{i=1}^K \left[ \min \left( \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}} A_i, \text{clip} \left( \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right]. \quad (7)$$

Here,  $\pi_{\theta_{\text{old}}}$  is the policy from the previous iteration,  $\pi_{\text{ref}}$  is the reference policy,  $\epsilon$  is the clipping coefficient for the policy ratio, and  $\beta$  controls the KL divergence penalty.

The key to GRPO is the reward design. Our goal is to ensure the model responds accurately while minimizing hallucinations. The output of the model will be defined as  $y$ .

**Think Format Reward:**  $y$  must meet the formatting: `<think>...</think><answer>...</answer>`. Therefore, format rewards,  $R_{f_t}$  are defined as:

$$R_{f_t}(y) = \begin{cases} 1, & \text{if } y \text{ matches the think format,} \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

**Evidence Format Reward** The model’s perception results are wrapped in a special format, which is required to be `<start="t1", end="t2", desc="description">`. Therefore, the evidence format reward function is:

$$R_{f_e}(y) = \begin{cases} 1, & \text{if } y \text{ matches the evidence format,} \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

**Accuracy Reward:** Our training data mainly contains five kinds of tasks: Video Temporal Grounding (VTG), Multiple Choice (MC), MC&VTG (GLUE), ReOrdering (RO), and Open-Ended (OE), illustrated in Fig. 3b. Except for the OE task, we use rule matching. And for the OE task, we use an open-source verification model (VM) [3]. Given the question  $Q$ , the ground truth answer  $G$ , and the model’s generated answer  $A_{\text{gen}}$  as input, we obtain the token probabilities  $P_{Ic}$  and  $P_C$  for the Incorrect and Correct labels, respectively. Thus, the accuracy reward  $R_{acc}$  is following:

$$R_a(y) = \begin{cases} IoU(G, y), & \text{if VTG} \\ Match(G, y), & \text{if MC} \\ \frac{P_C}{P_{Ic} + P_C}, & \text{if OE} \\ Match(G, y) + IoU(G, y), & \text{else} \end{cases} \quad (10)$$

where *Match* is defined as choosing the same options or having the same sorting order.

**Anti-hallucination Reward:** In sec. 3, Factual-Aware Evaluator( $E$ ) is trained for anti-hallucination reward. The reasoning model will output multiple visual evidences to describe video clips. Each evidence contains three pieces of information:  $t_{start}$ ,  $t_{end}$ , and  $desc$ . The probability that  $E$  judges whether the description is correct is:

$$P_{Y/N} = E(\text{clip}(Video, t_{start}, t_{end}), desc), \quad (11)$$

where `clip` is defined as using `ffmpeg` to obtain video clips, and  $Y$  means the description is correct. We also need to consider the quantity of evidence and ensure that different pieces of evidence do not describe the same content. For this purpose, the attenuation coefficient  $w$  is defined as the product of the maximum IoU and ROUGE between the current evidence and other evidence:

$$w_i = 1 - \max_{j \neq i} (IoU(e_i, e_j) \times ROUGE(e_i, e_j)). \quad (12)$$

Thus, our anti-hallucination reward ( $R_h$ ) is defined as a weighted sum of evidence scores, normalized by a non-linear function of  $n$ :

$$R_h(y) = \frac{1}{\max(0.6 + 0.8 \times n, n)} \sum_{i=1}^n \frac{w_i P_{Y,i}}{P_{N,i} + P_{Y,i}} \quad (13)$$

$R_h$  is only provided for the output where  $R_a$  is greater than 0.5. Considering that the  $R_a$  for the VTG and GLUE tasks is dense,  $R_h$  may affect the IoU results.

The overall reward  $R$  for a sample is computed as follows:

$$R = R_a + 0.5R_{f_t} + 0.5R_{f_e} + 0.2R_h \quad (14)$$

## 5. Experiment

### 5.1. Setups

**Training Details** For the Factual-Aware Evaluator, we use LoRA [20] based on Qwen2.5VL 7B, training for one epoch with a learning rate of 1e-4, where  $\lambda$  in formula 7 is 0.5, with a maximum of 16 frames and the total number of pixels limited to 12002828. For the reasoning model, training includes a cold-start SFT and RL using GRPO. The cold-start data incorporates 20K image CoT data from the Reason-RFT-CoT-Dataset [30] and our VideoPLR-14K. We collected 6K questions for training, sourced from LLaVA-Video-178K [45], Time-R1 [36], Cinepile [28], NextGQA [38], and ANet [10], as show in Fig. 3b. Both SFT and RL, the learning rate is set to 2e-6, and train for one epoch. We set the number of generations to 8 and the batch size to 6, the  $\beta$  eq. 7 is set to 0.01, and

Table 1. The results of Video Reasoning and General Benchmarks. For VCR-Bench, Video-tt, and TVG-Bench, we re-tested the performance metrics of each model. The metric of TVG-Bench is R1@0.5. Considering the fairness, for Videorchat-R1.5, we limited the maximum number of frames to 64, the maximum number of pixels to  $6000 \times 28 \times 28$ , and used single inference mode. The other models were evaluated using the default settings in the paper. In the Training Data column, we compared the data scales in the cold start and RL phases, in the form of cold start + RL phase.

Model	Training Data ↓	Reasoning					General	Temporal	Avg
		Video-Holmes	MMVU(mc)	VCR-Bench	VSI-Bench	Video-tt(mc)	VideoMME	TVG-Bench	
<b>Closed-Source Model</b>									
GPT-4o [9]	-	42.0	67.4	46.9	34.0	46.6	71.9	-	-
Gemini-2.0-Flash [7]	-	30.6	65.9	51.7	45.4	-	-	-	-
<b>3B Open-Source Model</b>									
Qwen2.5VL [2]	-	17.0	53.0	24.8	28.6	36.9	52.1	14.3	32.4
Time-R1 [36]	0+2.5K	32.2	52.5	22.9	28.5	34.9	53.4	<u>21.0</u>	35.3
Video-RFT [32]	102+8K	<u>34.3</u>	<u>52.6</u>	<u>25.8</u>	<u>31.9</u>	<u>35.2</u>	<u>54.7</u>	10.6	<u>34.7</u>
Video-PLR (our)	38+6K	<b>39.8</b>	<b>59.3</b>	<b>28.4</b>	<b>35.1</b>	<b>40.9</b>	<b>57.3</b>	<b>24.9</b>	<b>40.8</b>
<b>7B Open-Source Model</b>									
Qwen2.5VL [2]	-	27.8	63.3	29.7	31.8	39.9	60.1	18.7	38.8
Time-R1 [36]	0+2.5K	37.1	64.2	29.6	28.4	39.0	60.9	29.6	41.3
Video-R1 [12]	165+8K	36.5	64.2	<u>31.3</u>	35.8	39.1	59.3	13.7	40.0
Video-RFT [32]	102+8K	<u>39.8</u>	<b>68.5</b>	30.6	36.8	40.3	59.8	15.4	41.6
Videochat-R1 [23]	0+18K	33.0	63.8	30.0	31.3	39.1	60.8	30.1	41.2
Videochat-R1.5 [40]	0+80K	38.5	64.5	30.0	<b>40.6</b>	<u>41.3</u>	<u>61.2</u>	26.8	<u>43.2</u>
Video-PLR (our)	34+6K	<b>42.6</b>	<u>64.7</u>	<b>32.4</b>	<u>38.3</u>	<b>42.7</b>	<b>61.8</b>	<b>31.7</b>	<b>44.9</b>

Table 2. The results of hallucination benchmarks. † is defined as rewriting questions as captions. ‘Diff’ refers to the difference in accuracy between the correct and incorrect categories. Compared with Qwen2.5VL, our FAE has better performance and achieves capabilities at the level of closed-source models.

Model	VideoHalluciner [34]		HEAVEN† [14]		VidHalluc [22]
	Acc↑	Diff↓	Acc↑	Diff↓	MCQ Acc ↑
GPT-4o [9]	74.6	<b>0.9</b>	-	-	<b>91.0</b>
Gemini-1.5-Pro [8]	62.9	41.3	-	-	79.2
Qwen2.5VL [2]	67.8	22.4	63.7	28.9	89.2
FAE(ours)	<b>75.4 (+7.5)</b>	7.3 (-15.1)	<b>72.9 (+9.2)</b>	<b>2.1 (-26.8)</b>	90.3 (+1.1)

the maximum output length is set to 1600 tokens. The maximum number of frames and total pixels are set to 64 and  $2048 \times 28 \times 28$ . Our training code is optimized based on ms-swift [47] to adapt to the anti-hallucination reward, which incurs significant computational latency. Thus, we implement GRPO training without additional time consumption. **Evaluation Details** First, we evaluate the Factual-Aware Evaluator, including the VidHalluc [22], VideoHalluciner [34], and HEAVEN [14]. Set the maximum number of frames to 16, the number of pixels to  $2400 \times 28 \times 28$ , and the temperature to 0. For video reasoning models, the evaluation benchmarks include three types: reasoning, general, and temporal. The reasoning benchmarks include: Video-Holmes [6], MMVU [48], VCR-benchmark [27], Video-tt [46], VSI-bench [41]. The general benchmark is VideoMME [13], and the temporal benchmark is TVG-Bench [36]. Set the maximum number of

frames to 64, the number of pixels to  $6000 \times 28 \times 28$ , and the temperature to 0.

## 5.2. Hallucination Evaluation

The evaluation begins with the hallucination benchmark evaluation, as shown in Tab 2. We use VideoHalluciner [34] for binary judgment, the open-source HEAVEN [14] data reformatted as captions, and the MCQ task from VidHalluc [22]. On VideoHalluciner, FAE surpasses GPT-4o. It boosts accuracy over the base model by 7.5% and 9.2%. FAE improves judgment fairness, reducing bias towards ‘Y’ or ‘N’. It narrows the accuracy gap between correct and incorrect classifications by 15.1% and 26.8%. This reduction validates our data construction method and FAE’s reliability as an anti-hallucination reward, confirming the robustness and successful transfer of its learned abilities. Despite being trained exclusively on binary classification tasks, our model demonstrates strong generalization to the MCQ task. This indicates that the learned ability to judge hallucinations is robust and successfully transferred, validating the effectiveness of our training approach.

## 5.3. Main Results

We compared the models that were post-trained based on Qwen2.5VL. These baselines have different focuses in training. Among these models, Time-R1 and VideoChatR1 pay more attention to time perception, while VideoR1 and VideoRFT focus more on reasoning, and VideoChatR1.5 takes both into account. Our Video-PLR has the best per-

Table 3. Ablation study on the training data. It can be found that the combination of CoT and direct output can significantly enhance the hallucination evaluation ability.

Training Set		HAVEN <sup>†</sup> [14]		VideoHalluciner [34]		
Direct	CoT	Acc <sup>↑</sup>	Diff <sup>↑</sup>	Acc <sup>↑</sup>	Diff <sup>↓</sup>	Gather <sup>↑</sup>
–	–	63.7	28.9	67.8	22.4	41.7
–	✓	<u>72.6</u>	<b>0.6</b>	71.6	<b>6.3</b>	<u>51.6</u>
✓	–	70.1	4.6	<u>72.5</u>	28.8	47.5
✓	✓	<b>72.9</b>	<u>2.1</u>	<b>75.3</b>	<u>7.3</u>	<b>54.2</b>

formance and data efficiency at the same time.

**SOTA Results:** Our Video-PLR achieves a significant average improvement in both the 3B and 7B parameter scales, as shown in Tab. 1. In the comparisons with Video-Holmes, VCR-Bench, and Video-tt, our Video-PLR shows a stable improvement compared to VideoRFT and Video-R1 in terms of parameter scales of 3B and 7B. These evaluation benchmarks have more multi-hop problems, that is, the answers are distributed over multiple time periods and require joint reasoning. Our PLR paradigm has more advantages in handling such tasks. MMVU focuses more on scientific questions, and VideoRFT with 7B has the best performance, which may be because there is more domain knowledge in its training data. VideoChatR1.5 performs best in VSI-Bench. This benchmark requires a stronger spatial reasoning ability, and the training of VideoChatR1.5 includes a large number of image grounding and video tracking tasks, meeting this requirement. Surprisingly, our Video-PLR performs outstandingly in TVG-Bench. Time-R1 and VideoChat-R1 have trained a large number of VTG tasks, but do not have better performance than our model. This is related to their lack of a logic-rich cold start step and only optimizing the ability to output timestamps. The evaluation result of Video-PLR for the general benchmark, VideoMME, is also the best.

**Data Efficient:** While achieving the best performance, Video-PLR demonstrates excellent data efficiency. In the comparison at the 7B scale, our model only used a total of 34K cold starts and 6K for RL. The sampling for RL was 8 times in all cases, which can be approximately equivalent to the training consumption of 8 times the cold start. So, except for Time-R1, all the other models used far more training consumption than ours. Video-RFT and VideoChatR1 are approximately 2 times ours, Video-R1 is about 3 times ours, and VideoChat-R1.5 is close to 8 times ours. It should be noted that although we only used 0.5K of VTG training data, the effect has exceeded that of TimeR1 trained with 2.5K data in TVG-Bench.

## 5.4. Ablation Study

**Hallucination Training Data** As shown in Tab. 3, the mixed training data helps the model achieve better perfor-

Table 4. Ablate cold-start data and the hallucination reward design. We implement the same training configuration. Training with the anti-hallucination reward achieves the best performance.

Cold Start	Rewards	Video-Holmes	MMVU	VCR
w/o sft	Only Acc	35.1	55.6	25.4
Video-R1(165K)	Only Acc	37.8	58.6	27.2
Video-RFT(102K)	Only Acc	37.6	54.4	27.6
	+ Anti-H	37.0	58.2	<b>28.5</b>
our w/ PLR(34K)	Only Acc	36.8	58.5	26.6
	+ Anti-H w/o norm	36.9	56.0	26.2
	+ Anti-H	<b>39.8</b>	<b>59.3</b>	28.4

mance in hallucination evaluation. Specifically, we provide the accuracy difference for the binary classification task. The addition of CoT data has significantly alleviated the classification bias. This proves the effectiveness of the CoT data design, forcing the model to answer based on the video content. Additionally, directly answering ‘Yes’ or ‘No’ is demonstrably more efficient than answering with CoT. The performance of the model trained with a mixture is also better than that trained only with CoT. The Gather column shows the comprehensive performance, which requires both answers to paired questions to be correct. Obviously, mixed training is superior to other solutions. Thus, Hybrid training yields more efficient responses and more accurate hallucination judgments.

**Training Strategies** As shown in Tab. 5, all the settings using cold start are better than those without cold start. Compared with Video-R1 and Video-RFT, the larger scale of the cold-start data brings better performance with only an accuracy reward. When adding the anti-hallucination reward (+Anti-H), our model achieved a more stable improvement compared to Video-RFT. This stems from the fact that the anti-hallucination at the segment level is more reliable, and the reliable perception stimulates the effectiveness of the PLR paradigm. Then we compare with the non-linear norm design of the anti-hallucination reward (eq. 13). The non-linear norm encourages the model to output more visual perception, because this design provides a lower reward for less visual evidence. When using the mean score of all evidence as the reward, the model tends to be conservative and only outputs reliable perceptual results, which creates a lack of evidence and leads to reduced performance. It should be noted that adding the anti-hallucination reward leads to a general improvement in the model, demonstrating the effectiveness of anti-hallucination reward design.

Supplementary materials provide more discussion, statistical data, training curves, and so on.

## 6. Conclusion

In this paper, we present Perception Loop Reasoning (PLR), a novel video reasoning paradigm that iteratively perceives and reasons about videos, ensuring sufficient

visual evidence. To further ensure the reliability of visual perception, the Factual-Aware Evaluator participates in the reinforcement learning stage as an anti-hallucination reward, which encourages the model to provide sufficient and precise visual evidence. The hallucination preference data we carefully constructed enable the FAE to have reliable description-video consistency judgment, reaching the level of GPT-4o. Our Video-PLR achieves SOTA on both 3B and 7B parameter scales and is more data-efficient compared to previous methods. The PLR paradigm contributes a new implementation idea for video reasoning and proposes that perception is the basis of reasoning. We hope our contributions can promote the further development of video reasoning with an emphasis on the reliability of the reasoning process.

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 3
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. 7
- [3] Ding Chen, Qingchen Yu, Pengyuan Wang, Wentao Zhang, Bo Tang, Feiyu Xiong, Xinchu Li, Minchuan Yang, and Zhiyu Li. xverify: Efficient answer verifier for reasoning model evaluations. *arXiv preprint arXiv:2504.10481*, 2025. 6
- [4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1
- [5] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1
- [6] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning?, 2025. 7
- [7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasapat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7, 1
- [8] Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 7, 1
- [9] OpenAI et al. Gpt-4o system card, 2024. 7
- [10] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2, 3, 5, 6
- [11] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*, 2025. 2
- [12] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv e-prints*, pages arXiv–2503, 2025. 1, 3, 7
- [13] Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 7
- [14] Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation, 2025. 7, 8
- [15] Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, Jinwen Luo, Weibo Gu, Zexuan Li, Xiaojing Zhang, Yangyu Tao, Han Hu, Di Wang, and Ying Shan. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts, 2025. 3
- [16] Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning. *arXiv preprint arXiv:2506.00318*, 2025. 2
- [17] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 1, 2, 3
- [18] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191, 2025. 2
- [19] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024. 4

- [20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6
- [21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer, 2024. 1
- [22] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13723–13733, 2025. 2, 3, 7
- [23] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 3, 7
- [24] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*, 2025. 3
- [25] OpenAI. OpenAI o1 System Card. <https://openai.com/index/openai-o1-system-card/>, 2024. Accessed: 2025-10-20. 2
- [26] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *arXiv preprint arXiv:2506.07464*, 2025. 3
- [27] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning, 2025. 7
- [28] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark, 2024. 6
- [29] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3, 6
- [30] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025. 6
- [31] Michael Tschanen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 3
- [32] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorf1t: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. 1, 3, 7
- [33] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, pages 396–416. Springer, 2024. 1
- [34] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohalluc: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024. 2, 3, 7, 8
- [35] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 92–101, 2024. 2
- [36] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025. 3, 6, 7
- [37] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 5
- [38] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 6
- [39] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *CoRR*, 2024. 2
- [40] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025. 3, 7
- [41] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 7
- [42] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2
- [43] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller llms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025. 1
- [44] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 1
- [45] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *CoRR*, 2024. 1, 6
- [46] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic

benchmark for advanced video reasoning and understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20626–20636, 2025. [7](#)

- [47] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift: a scalable lightweight infrastructure for fine-tuning, 2024. [7](#)
- [48] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding, 2025. [7](#)
- [49] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. [1](#)

# Alternating Perception-Reasoning for Hallucination-Resistant Video Understanding

## Supplementary Material

### Overview

The supplementary material is organized as follows:

- **Sec. 7** discusses the impact of the anti-hallucination reward and outlines current limitations.
- **Sec. 8** presents additional experiments on reasoning honesty, training visualization, and inference efficiency analysis.
- **Sec. 9** details training hyperparameters and our parallel reward optimization strategy for efficiency.
- **Sec. 10** provides a comprehensive statistical analysis and bias quantification for the training datasets.
- **Sec. 11, 12, 13** showcase qualitative data examples, detailed case studies, and data generation prompts.

### 7. Discussion

The anti-hallucination reward is one of the main contributions. In the main manuscript, the non-linear norm design and the anti-hallucination reward significantly affect the performance of the model, as illustrated in the ablation study. Sec. 8.2 further illustrates the trend changes in model outputs during training under different settings. The primary metrics observed are the changes in model output length and the quantity of evidence enumerated. The simplistic design of averaging anti-hallucination rewards has led to a decline in both output length and the number of perceptions, resulting in performance loss. This further supports that comprehensive perceptual evidence makes video reasoning more reliable and enhances performance. Also, the standard anti-hallucination encourages the model perception, resulting in an increase in both output length and the number of perceptions. Moreover, the crucial role of anti-hallucination reward is the Factual-Aware Evaluator (FAE). Although the hallucination benchmarks show the effectiveness of the FAE, the FAE is used with its probability. More detailed analysis and discussion are necessary to verify that FAE is superior to the baseline. Therefore, Sec. 8.5 presents the reward margins of different models when used as reward models, where a larger margin indicates a clearer reward signal.

**Limitations.** First, the PLR paradigm relies on the precise temporal grounding capabilities of the underlying VideoLLM. While some open-source models [4, 5, 44, 49] currently face challenges in timestamp generation, we anticipate that rapid advancements in foundation models (e.g., Gemini series [7, 8]) will alleviate this bottleneck. Second, due to computational constraints, we restricted our

experiments to 3B and 7B parameter scales with efficient data sizes (VideoPLR-14K and AnetHallu-117K) and controlled resolutions, rather than training larger-scale (e.g., 72B) exhaustive models. Despite these constraints, VideoPLR achieves state-of-the-art performance and high data efficiency, confirming the validity of our paradigm.

### 8. More Reasoning Results

#### 8.1. Honest Reasoning

Through an analysis of the training logs, we observed instances of unfaithful reasoning, a phenomenon also noted in recent literature [1]. Manifestations of such unfaithfulness include prematurely terminating the perception loop despite explicitly identifying a need for further evidence, as well as discrepancies between the reasoning chain’s conclusion and the final output answer. We posit that accurate perception is the prerequisite for sound logic; therefore, the primary objective of this work is to ensure the reliability of visual perception rather than the perfection of textual logic. Although optimizing textual logic falls outside the primary scope of this study, we have nonetheless included relevant experiments to address this aspect, as detailed in Tab. 5. Specifically, we employed Qwen2.5-32B-AWQ as a judge to evaluate the consistency between the generated reasoning and the final answer. The prompt instructs the model to assess solely whether the output is self-consistent, disregarding the presence of logical fallacies. The probability of the “Match” token is then utilized as the reward signal.

However, experimental results indicate that employing an off-the-shelf language model as a judge, even at the 32B parameter scale, fails to enhance overall performance. As shown in the last row of Tab. 5, the addition of the consistency reward (+ *Honest*) leads to a performance decline compared to the standard setting. While we observed a reduction in unfaithful behaviors during training, this did not translate into benchmark improvements. We hypothesize that this is due to the inherent preference bias of the non-fine-tuned judge model. The policy model likely overfitted to these specific preferences, resulting in rigid output patterns that hindered the reasoning flexibility required for complex video understanding tasks.

#### 8.2. Training Visualization

Thinking length and the number of perceptions are key indicators of model behavior. We find that the standard anti-hallucination reward encourages the model to gener-

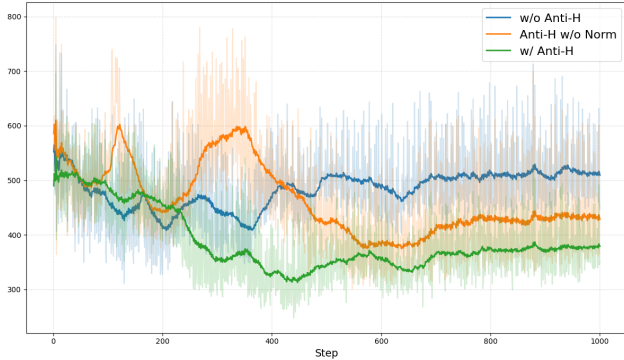


Figure 5. This figure records the trend of changes in the number of tokens output by the model as training steps progress.

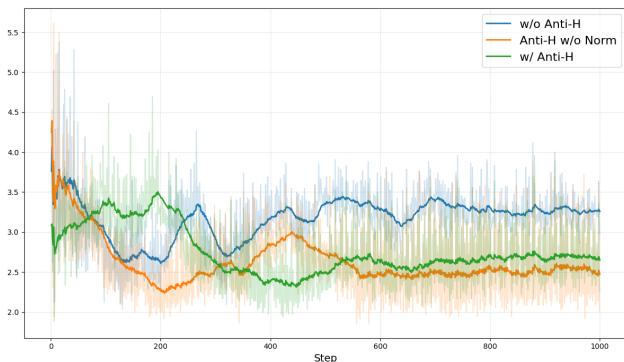


Figure 6. This figure records the trend of changes in the number of perceptions within the thinking by the model as training steps progress.

Table 5. Ablation Study Table from the main manuscript.

Cold Start	Rewards	Video-Holmes	MMVU	VCR
our w/ PLR(34K) Only	Acc	36.8	58.5	26.6
+ Anti-H w/o norm		36.9	56.0	26.2
+ Anti-H		<b>39.8</b>	<b>59.3</b>	<b>28.4</b>
+ Anti-H +Honest		35.0	55.3	26.4

ate more sufficient perceptions. Compared to the baseline without the anti-hallucination reward, the thinking length increases significantly, as shown in Fig. 5. This is attributed to the non-linear normalization strategy. Specifically, the reward degrades to a mean calculation when the perception count is at least three; however, for fewer perceptions, the reward is lower than the mean, acting as a penalty. Using the no-reward setting as a baseline, we observe that employing a simple mean reward increases thinking length but decreases the number of perceptions, as shown in Fig. 6. Tab. 5 indicates that the significant performance improvement is driven by the increased quantity of perceptions. The performance drop observed with longer outputs after

removing the normalization further confirms that thinking length alone is not the key factor for performance gain.

Table 6. nextgqa

Model	NextGQA	
	Acc	mIoU
Qwen2.5-VL-7B	59.5	15.4
Videorchat-R1	70.6	32.4
Videorchat-R1-Thinking	69.2	<b>36.1</b>
Video-PLR (our)	<b>77.9</b>	31.1

### 8.3. NextGQA

Our RL training data includes 1k qa from NextGQA. Therefore, we evaluate our 7B Video-PLR and compare it with VideoChatR1, which is tuned with the full set of NextGQA. As shown in Tab. 6, VideoPLR surpasses all baselines in terms of accuracy. Specifically, it outperforms the Qwen2.5-VL-7B baseline by 18.4% and leads VideoChat-R1 by 7.3%. While VideoChat-R1-Thinking shows a slight advantage in mIoU, our model achieves the best overall answer accuracy with significantly less training data.

### 8.4. Reasoning Analysis

We evaluate the inference efficiency using VLLM on a single NVIDIA A40 GPU. As shown in Table 7, Video-PLR requires an average of 16.34s per query. While this latency is higher than the single-step baselines (e.g., 10.57s for Video-R1), it is a necessary trade-off for our dynamic and grounded reasoning, which is justified by the following observations.

First, regarding dynamic adaptability, shown in Fig. 7, Video-R1 exhibits a narrow distribution (100-500 tokens), suggesting a reliance on static, template-like responses. In contrast, Video-PLR shows a significantly broader dynamic range (300-1300 tokens). This indicates that our model adaptively adjusts its reasoning depth—allocating fewer tokens for simple queries and expanding its reasoning chain for complex tasks. Furthermore, regarding temporal awareness (Fig. 8), Video-PLR demonstrates a clear positive trend where both token count and evidence count increase with video length. This proves that the model is actively processing the temporal extent of the video rather than generating hallucinations of fixed length regardless of input complexity. Most importantly, as shown in Fig. 9, the increased computational cost translates directly into effective computation. There is a strong positive correlation between the generated token count and the number of valid evidence segments found. This confirms that the additional inference time is effectively utilized for the **Perception Loop** to gather necessary visual evidence, ensuring that the longer reasoning paths are grounded in facts.

Table 7. We report the average token length, token distribution range, and average inference latency per sample. All models are evaluated using VLLM on a single NVIDIA A40 GPU

Model	Avg. Tokens	Avg. Time (s)
Video-R1	298.13	10.57
Video-RFT	314.54	11.46
Video-PLR (Ours)	530.70	16.94

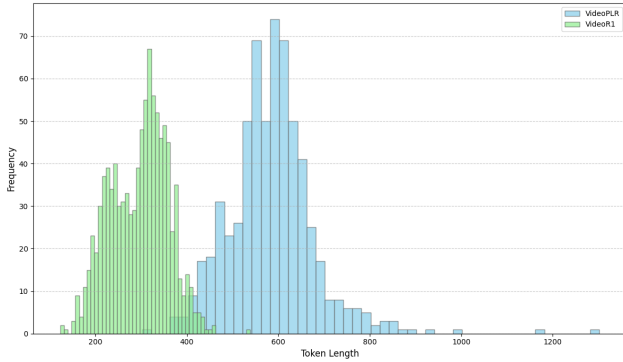


Figure 7. Compare the number of tokens output by Video R1 and VideoPLR in MMVU

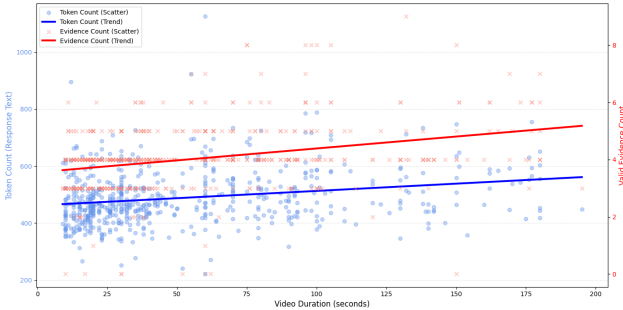


Figure 8. Statistics of the number of tokens output by the model and the amount of evidence provided for videos of different lengths

### 8.5. Reward Margin Analysis

To validate the effectiveness of our FAE as a reward model, we evaluate its discriminative capability on the HAVEN dataset compared to baselines (SigLip and Qwen2.5-VL). As presented in Tab. 8, our model achieves a significantly higher AUC of 76.32%, surpassing Qwen2.5-VL and SigLip by 7.7% and 15.9%, respectively.

More importantly, we analyze the reward margin—the distinctiveness between correct and incorrect captions. As illustrated in the histograms (Fig. 12 vs. Fig. 11 and Fig. 10), our model exhibits a much sharper decision boundary compared to the baselines. The separation gap ( $\Delta$ ) between the mean confidence for factual captions ( $P_{yes}$ ) and

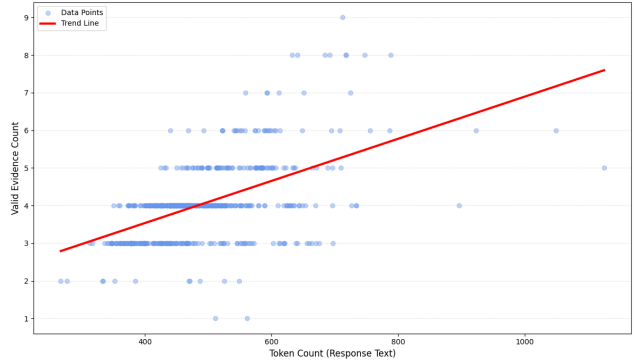


Figure 9. The statistical analysis of the relationship between the number of tokens output by the model and the amount of evidence included shows a positive correlation.

hallucinations ( $P_{no}$ ) reaches 29.62%, which is nearly  $3\times$  that of SigLip (11.32%). This quantitative gap demonstrates that our FAE provides a more definitive and high-quality reward signal, effectively encouraging the policy to distinguish between hallucinations and facts during RL training, whereas baselines suffer from ambiguous boundaries that may confuse the policy.

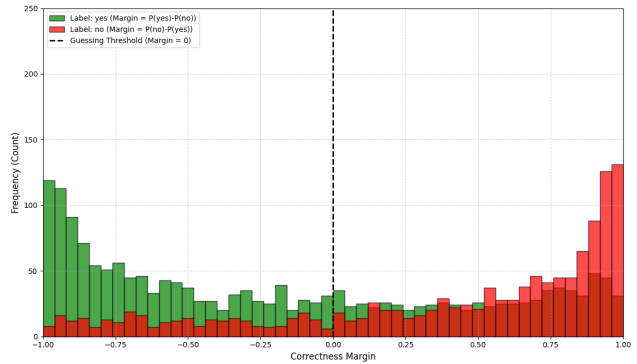


Figure 10. Prediction margin distribution of SigLip.

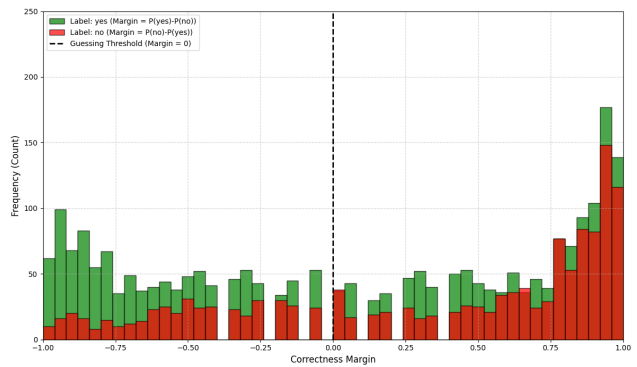


Figure 11. Prediction margin distribution of Qwen2.5-VL.

Table 8. **Quantitative comparison of reward distinctiveness on the HAVEN dataset.** We report the AUC, the mean reward for factual ( $R_{yes}$ ) and hallucinatory ( $R_{no}$ ) captions, and the separation gap ( $\Delta$ ). Our FAE provides the clearest decision boundary, offering a robust signal for RL training.

Model	AUC	Confidence Scores		
		Mean $R(\text{yes}) \uparrow$	Mean $R(\text{no}) \downarrow$	$\Delta$ (Gap) $\uparrow$
SigLip	60.46	40.96	<b>29.64</b>	11.32
Qwen2.5-VL	68.59	54.34	33.07	21.27
<b>FAE (Ours)</b>	<b>76.32</b>	<b>65.44</b>	35.82	<b>29.62</b>

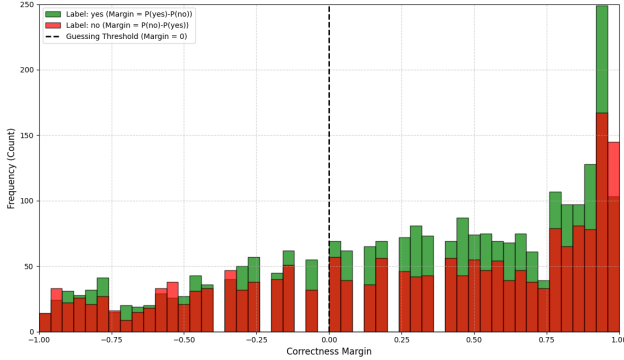


Figure 12. Prediction margin distribution of our FAE.

## 9. Training Details

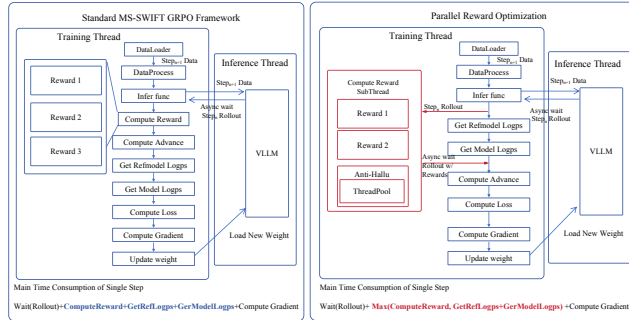


Figure 13. Comparison of the Standard MS-SWIFT GRPO Framework (left) and our Parallel Reward Optimization (right). In the standard framework, reward computation is a blocking operation. Our optimization decouples the reward computation into an asynchronous sub-thread. This allows the heavy inference of the Factual-Aware Evaluator to execute in parallel with the reference and policy model log-probability computations, effectively masking the external reward latency.

### 9.1. Training Efficiency Optimization

**Efficiency Optimization for GRPO Training.** Integrating the Factual-Aware Evaluator (FAE) as a reward model introduces significant computational overhead, as it requires

additional inference passes for every generated sample. In the standard MS-SWIFT GRPO framework, the Compute Reward stage is sequential, meaning the training thread must idly wait for all reward models to return results before proceeding to calculating log probabilities (Get Refmodel Logps and Get Model Logps). This serial execution leads to a training bottleneck:

$$T_{step} = T_{rollout} + T_{reward} + T_{logps} + T_{grad} \quad (15)$$

To address this, we implemented a **Parallel Reward Optimization** strategy. As illustrated in the right panel of Fig. 13, we decouple the reward computation into an asynchronous sub-thread. Upon receiving the rollout data (Step\_n Rollout), the main training thread immediately proceeds to the GPU-intensive log-probability calculations. Simultaneously, the sub-thread dispatches requests to the reward models. The main thread only performs an asynchronous wait (Async wait) after the log-probability calculations are complete. Since the time required for Get Refmodel Logps and Get Model Logps typically exceeds or matches the inference time of the reward models, the latency of the FAE is effectively masked. The optimized time consumption becomes:

$$T_{step} = T_{rollout} + \max(T_{reward}, T_{logps}) + T_{grad} \quad (16)$$

**Deployment and Load Balancing.** To ensure the reward computation throughput matches the training speed, we deployed the reward models on a separate dedicated inference cluster consisting of four NVIDIA RTX 4090 GPUs. Specifically:

- The **Verify Model (VM)** for the Open-Ended task is deployed on **1 × RTX 4090**.
- The **Factual-Aware Evaluator (FAE)** is deployed on **3 × RTX 4090**. We implemented a load-balancing mechanism that distributes the concurrent verification requests from the training batch across these three GPUs, ensuring optimal resource utilization and minimal latency.

This architecture allows us to utilize a complex, model-based anti-hallucination reward without increasing the overall wall-clock time of the training process.

## 9.2. Parameter Details

We present the parameter details of RL training in detail, as shown in Tab. 9 and Tab. 10.

Table 9. Training Setting of RL on 3B model

Parameter	Value
DeepSpeed	Zero3
Learning Rate	$2 \times 10^{-6}$
Batch Size	6
num_generations	6
temperature	1.0
top_p	0.95
top_k	50
$\beta$	0.01
max_grad_norm	1
Base Model	Qwen2.5VL 3B
Ref Model	Qwen2.5VL 3B
Training Epochs	1
Max Frames	64
Max Pixels	1,505,280

Table 10. Training Setting of RL on 7B model

Parameter	Value
DeepSpeed	Zero3
Learning Rate	$2 \times 10^{-6}$
Batch Size	6
num_generations	8
temperature	1.0
top_p	0.95
top_k	50
$\beta$	0.01
max_grad_norm	1
Base Model	Qwen2.5VL 7B
Ref Model	Qwen2.5VL 7B
Training Epochs	1
Max Frames	64
Max Pixels	1,505,280

## 10. Dataset Details

### 10.1. Dataset Information

In this section, we provide a detailed statistical analysis of the training data, covering both the AnetHallu-117K used for the anti-hallucination reward and the VideoPLR-14K used for the cold start stage. To ensure the Factual-Aware Evaluator (FAE) can robustly identify various types of errors without bias, we constructed the AnetHallu-117K

dataset with a diverse distribution of hallucination categories. As shown in Fig. 14, the dataset covers five distinct hallucination types to ensure comprehensive evaluation capabilities. The distribution is relatively balanced, with Attribute Modification (21.8%), Detail Conflation (21.3%), Quantity Modification (21.0%), and Action Substitution (20.2%) each accounting for approximately 20% of the data. Additionally, Temporal Reordering accounts for 15.6%, focusing specifically on the chronological logic of video events. This diversity is crucial for preventing the model from overfitting to specific error patterns.

Beyond type diversity, we strictly controlled the label distribution to mitigate potential bias towards Yes or No answers. As illustrated in Fig. 15, the ratio of Positive (Yes) to Negative (No) samples is 46.6% to 53.4%, achieving a near-balanced state. To further enhance the model’s reasoning ability while maintaining inference efficiency, we employed a hybrid response format. As shown in Fig. 16, 34.1% of the data requires Chain-of-Thought (CoT) reasoning, while 65.9% consists of direct answers. We also addressed linguistic shortcuts by analyzing vocabulary distributions. The word clouds in Fig. 17 and Fig. 18 demonstrate a diverse and overlapping vocabulary across positive and negative captions. This confirms that our Text Bias Elimination pipeline effectively reduces linguistic cues that might otherwise lead to spurious correlations, ensuring the model judges based on visual content rather than word frequency.

For the cold start stage, we utilized the VideoPLR-14K dataset to initialize the Perception Loop Reasoning (PLR) paradigm. As shown in Fig. 19, this dataset is primarily constructed from two sources: NextQA (62.5%) and ActivityNet (37.5%). NextQA provides rich causal and temporal logic questions, which were re-annotated with timestamped reasoning chains to train the model’s sequential perception ability. Complementing this, ActivityNet serves as the source for the ReOrdering task, focusing on non-sequential event sorting and dense captioning to enhance the model’s global understanding of video structures. Finally, we analyzed the distribution of token lengths within the generated Chain-of-Thought (CoT) sequences. As presented in Fig. 15, the reasoning chains exhibit a substantial length, with a mean of 502.53 tokens and a median of 474.00 tokens. The distribution follows a natural pattern with a long tail extending beyond 1,200 tokens, indicating that the dataset encompasses a wide range of reasoning complexities. This length distribution confirms that our cold-start data provides rich and detailed reasoning contexts rather than superficial descriptions, thereby effectively initializing the model’s capability for deep perception loop reasoning.

### 10.2. Quantifying Textual Bias

To rigorously evaluate the effectiveness of our debiasing approach, we introduce the *Vocabulary Polarity Score* based

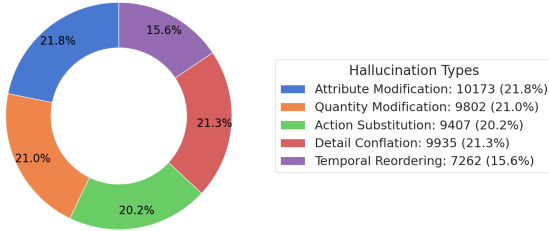


Figure 14. Distribution of the five hallucination types in the AnetHallu-117K dataset.

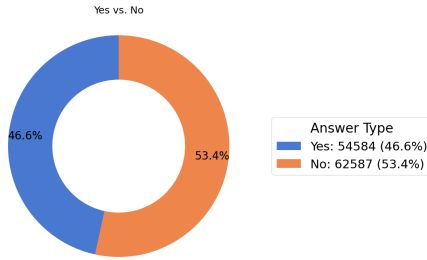


Figure 15. The balance statistics of Yes and No answers in the binary judgment tasks.

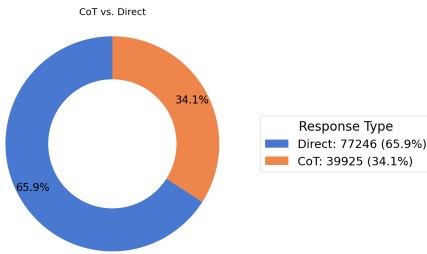


Figure 16. The ratio of Direct Answer versus Chain-of-Thought (CoT) reasoning in the mixed training data.

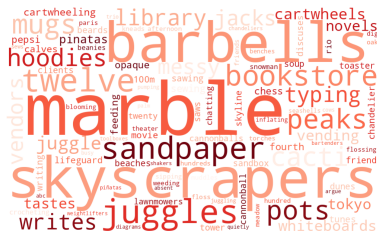


Figure 17. Word cloud visualization of high-frequency concepts in the negative (hallucinated) captions.

on the Log-Odds Ratio with Dirichlet smoothing. For a given word  $w$ , its polarity score is defined as:

$$\text{Score}(w) = \log \left( \frac{\hat{P}(w|\mathcal{D}_{pos})}{\hat{P}(w|\mathcal{D}_{neg})} \right) \quad (17)$$



Figure 18. Word cloud visualization of the vocabulary distribution in the positive (ground-truth) captions.

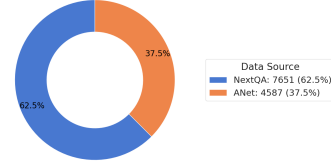


Figure 19. Source composition of the Cold-Start SFT data, consisting of NextQA and ActivityNet (ANet).

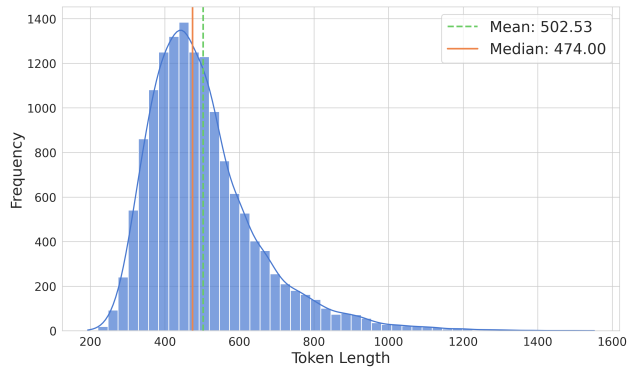


Figure 20. Distribution of token lengths in the generated reasoning chains (CoT).

where  $\hat{P}(w|\mathcal{D}_{pos})$  and  $\hat{P}(w|\mathcal{D}_{neg})$  represent the smoothed probabilities of word  $w$  appearing in the positive and negative caption sets, respectively. We utilize Dirichlet smoothing (adding a pseudo-count of 1) to prevent numerical instability caused by zero-frequency words.

A score significantly greater than 0 indicates a strong association with positive captions, while a score significantly less than 0 implies a negative bias. To quantify the overall bias of the dataset, we define the **Mean Absolute Polarization (MAP)**:

$$\text{MAP} = \frac{1}{|V|} \sum_{w \in V} |\text{Score}(w)| \quad (18)$$

where  $V$  denotes the vocabulary set. A lower MAP value indicates that the vocabulary distribution is more balanced, with fewer words serving as discriminative shortcuts for the model.

### 10.3. Analysis of Debiasing Effectiveness

Based on the proposed vocabulary polarity score, we design an iterative filtering mechanism to progressively neutralize the dataset. The detailed procedure is outlined in **Algorithm 1**.

We conducted a quantitative analysis to verify that our iterative filtering method reduces lexical cues without compromising semantic richness.

**Distribution Shift.** As visually presented in Figure 21, the original dataset (depicted in red) exhibits a dispersive distribution with heavy tails. This high variance suggests a significant presence of highly polarized words (extreme Log-Odds scores), which potentially serve as spurious correlations for the model. In contrast, the debiased dataset (green) demonstrates a distinct distributional shift towards a leptokurtic shape centered at zero. The contraction of the distribution tails confirms the successful removal of extreme lexical artifacts, resulting in a more neutral vocabulary profile.

**Iterative Convergence.** Furthermore, Figure 22 illustrates the dynamic optimization process of our method. The MAP score exhibits a monotonic downward trend over the course of 15 iterations. The steep decline in the initial phases (Iterations 0-6) corresponds to the elimination of the most obvious biased terms. As the process continues, the curve flattens, indicating that the algorithm progressively fine-tunes the dataset by addressing more subtle lexical imbalances. This convergence validates the stability of our penalty-based filtering strategy.

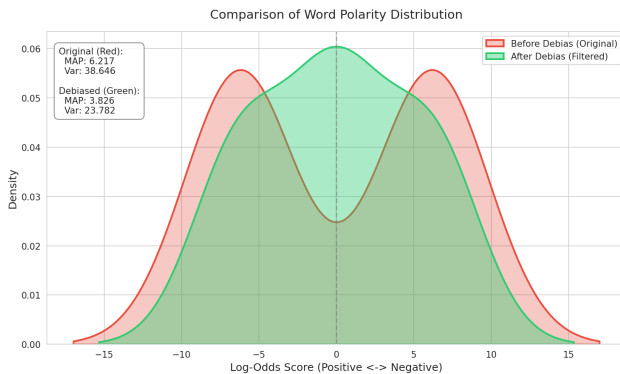


Figure 21. **Comparison of Vocabulary Polarity Distributions.** The plot contrasts the density of Log-Odds scores between the original dataset (Red) and the debiased dataset (Green).

### 11. Data Example

From Figure 23 to Figure 25, we present some cold-start data and hallucination preference data. The cold-start data is divided into two types: multi-choice and open-ended, as shown in Figure 23 and Figure 24, respectively.

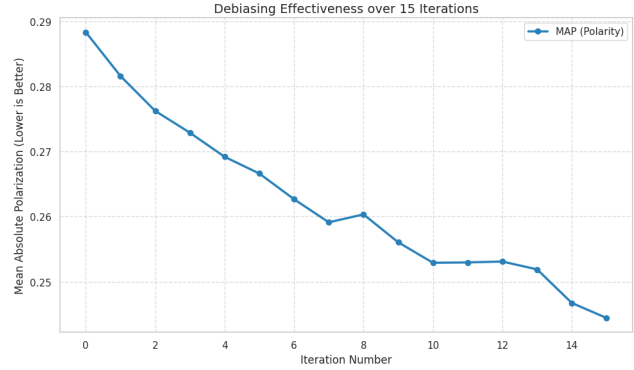


Figure 22. **Convergence of MAP Score.** The line graph tracks the reduction in Mean Absolute Polarization across 15 filtering iterations.

### 12. Case Study

Here we present cases from Video-Holmes (Figure 26) and VCR-Bench (Figure 27). Video-Holmes is a bench that requires strong video inference capabilities. In this bench, Video-PLR demonstrates powerful inference capabilities, effectively utilizing its perception loop mechanism to perceive implicit event relationships in the environment. In contrast, VideoRFT and VideoChat-R1.5 remain at the level of understanding the scene itself, failing to capture the implicit relationships between events. In VCR-Bench, Video-PLR’s perceptual loop mechanism reduces the omission of key content and improves the model’s overall grasp of the video.

### 13. Prompts

In this section, we provide the detailed prompts used in our data construction pipeline. Fig. 28 to Fig. 31 illustrate the four-step pipeline for synthesizing the **VideoPLR-14K** cold-start data. Fig. 32 to Fig. 35 display the prompts for constructing the **AnetHallu-117K** dataset, including ground-truth generation, hallucination injection, and correction. Finally, Fig. 36 presents the inference system prompt for Video-PLR.

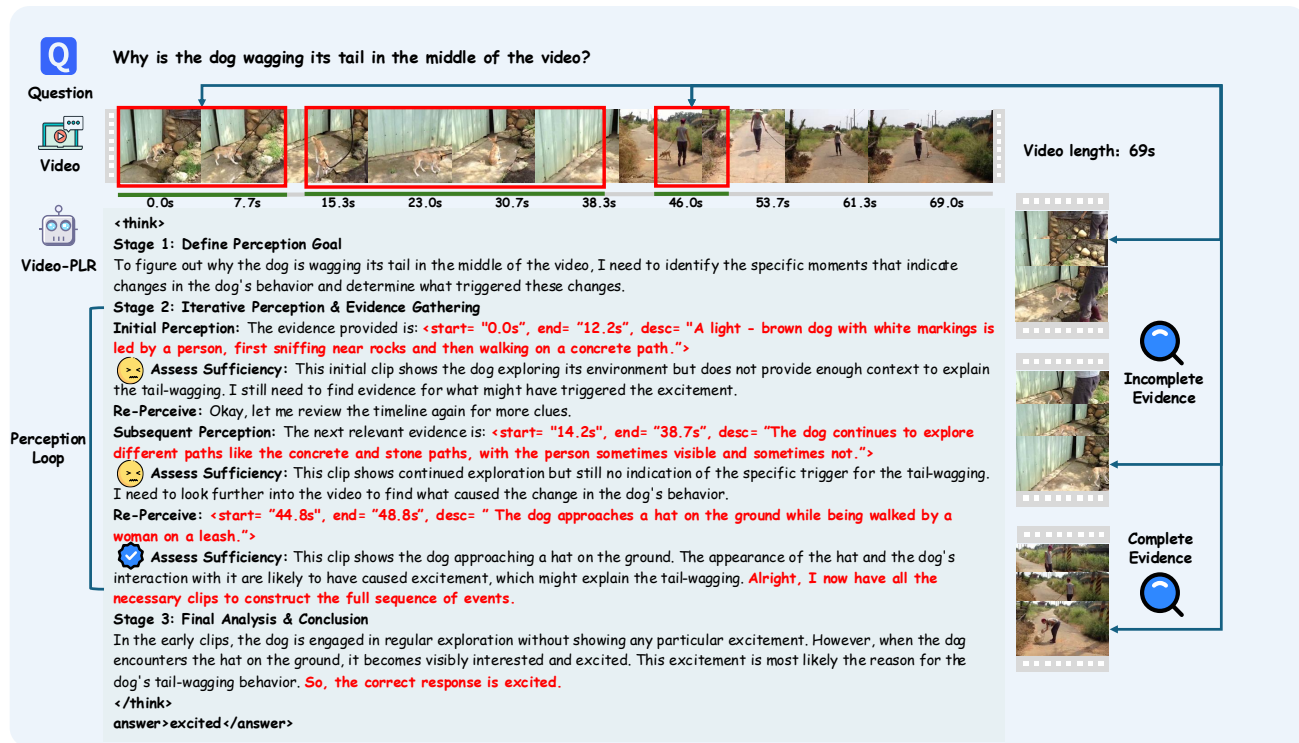


Figure 23. Visualization of the cold-start data in open-ended format.

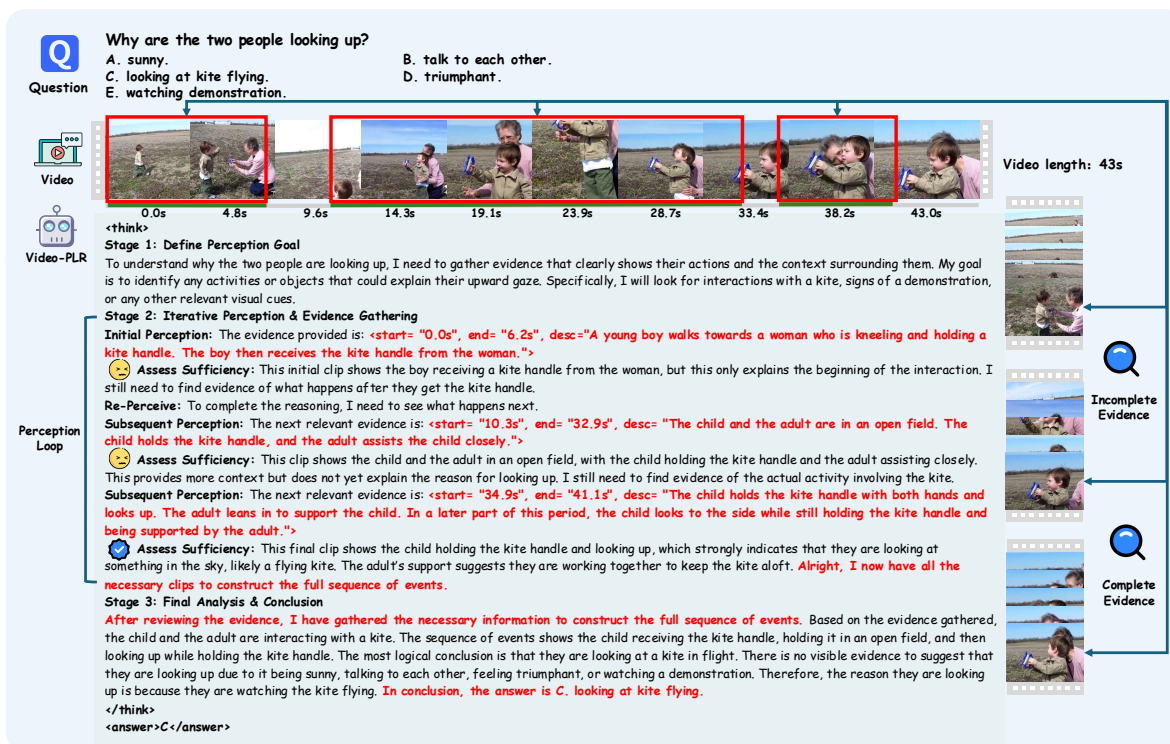


Figure 24. Visualization of the cold-start data in multi-choice format.

---

**Algorithm 1** Bias Eliminate Algorithm

---

**Input:**  $D_{in}$ : Full dataset of (pos, neg) pairs  
 $N_{iter}$ : Number of filter iterations  
 $P_p, P_n$ : Total filter percentage (pos, neg)

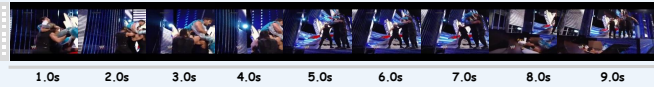
**Output:**  $D_{out}$ : The debiased dataset

- 1:  $D \leftarrow D_{in}$
- 2:  $\delta_p \leftarrow P_p/N_{iter}, \delta_n \leftarrow P_n/N_{iter}$   $\triangleright$  Per-iteration rate
- 3: **for**  $i \leftarrow 1$  to  $N_{iter}$  **do**
- 4:  $C_p, C_n \leftarrow \text{GETALLCAPTIONS}(D)$
- 5:  $M_p, M_n \leftarrow \text{CREATEPENALTYMAPS}(C_p, C_n)$   $\triangleright$   
Find bias
- 6:  $S_p \leftarrow \text{SCORECAPTIONS}(C_p, M_p)$   $\triangleright$  Score by bias
- 7:  $S_n \leftarrow \text{SCORECAPTIONS}(C_n, M_n)$
- 8:  $R_p \leftarrow \text{GETTOPPERCENT}(S_p, \delta_p)$   $\triangleright$  Removal set
- 9:  $R_n \leftarrow \text{GETTOPPERCENT}(S_n, \delta_n)$
- 10:  $D \leftarrow \text{FILTERDATASET}(D, R_p \cup R_n)$   $\triangleright$  Filter
- 11: **end for**
- 12:  $C_{surv} \leftarrow \text{GETALLCAPTIONS}(D)$   $\triangleright$  Get survivors
- 13:  $D_{out} \leftarrow \text{RECONSTRUCT}(D_{in}, C_{surv})$   $\triangleright$  Rebuild
- 14: **return**  $D_{out}$

---



Video



A series of **Positive and Negative** pairs to **Eliminate Model Bias** 😞



QA pairs

Analyze the video carefully and determine with certainty: Does the text: {desc} accurately describe what happens in the video?

**Attribute Modification**

**Positive:**  
 desc: Four people do a wrestling stunt on a red-and-blue elevated stage, two wear **blue tops** with white accents, two in black.  
 CoT: choose: The description is clear and accurate. The video depicts four individuals performing a wrestling stunt on an elevated stage with distinct clothing colors that match the description given. **The answer is Yes.**  
 reject: The action depicted involves four individuals engaging in a choreographed dance or group performance on an elevated stage featuring red and blue lighting effects. The participants' attire includes some with blue tops with white accents and others in black. **The answer is No.**

**Negative:**  
 desc: Four people do a wrestling stunt on a red-and-blue elevated stage; two wear **green tops** with white accents, two in black.  
 CoT: choose: The video shows four people on a stage, but it does not indicate that they are doing a wrestling stunt. They appear to be engaged in some form of performance or dance routine, not a wrestling stunt. The individuals involved include two wearing green tops with white accents and two in black outfits. **The answer is No.**  
 reject: The video shows four people, including two wearing green tops with white accents, performing a wrestling stunt on a red-and-blue elevated stage. Therefore, the description is accurate. **The answer is Yes.**

**Quantity Modification**

**Positive:**  
 desc: **Four individuals** perform a wrestling stunt on a red-and-blue elevated stage in front of a seated audience.  
 CoT: choose: The description accurately mentions four individuals performing a wrestling stunt on a red-and-blue elevated stage, although it doesn't specify the presence of a seated audience. Therefore, the main elements match those described in the text. **The answer is Yes.**  
 reject: There are four individuals performing a wrestling stunt on an elevated stage; however, the stage is not described as red and blue, and the audience is not mentioned. The description also does not mention that they are standing on a raised platform in front of a seated audience. **The answer is No.**

**Negative:**  
 desc: **Six individuals** perform a wrestling stunt on a red-and-blue elevated stage in front of a seated audience.  
 CoT: choose: The video shows individuals on a stage performs a stunt, but the stage is not red and blue elevated nor the context suggests an audience presence. Therefore, the description does not fully match the video content. **The answer is No.**  
 reject: The video indeed shows six individuals performing a wrestling stunt on a red-and-blue elevated stage in front of a seated audience. **The answer is Yes.**

**Action Substitution**

**Positive:**  
 desc: Four individuals perform a **wrestling stunt** on a brightly lit, elevated stage with a seated audience.  
 CoT: choose: The video shows four individuals performing a wrestling stunt on a brightly lit, elevated stage with a seated audience. The text accurately describes this scene. **The answer is Yes.**  
 reject: The response provided describes an incorrect action because it characterizes four individuals performing a wrestling stunt when the actual scenario depicts a dance performance by multiple dancers on a brightly lit stage. **The answer is No.**

**Negative:**  
 desc: Four individuals perform a **gymnastics routine** on a brightly lit, elevated stage with a seated audience.  
 CoT: choose: The video shows multiple individuals on a brightly lit stage, engaging in a kind of wrestling or combat performance rather than a gymnastics routine, and they do not include a seated audience. **The answer is No.**  
 reject: The video features four individuals performing a gymnastics routine on a brightly lit, elevated stage with a seated audience. Therefore, the description provided is accurate. **The answer is Yes.**

**Detail Conflation**

**Positive:**  
 desc: Four people do a wrestling stunt on a red-and-blue elevated concert stage with a **seated crowd** in front.  
 CoT: choose: The video does show four individuals engaging in a wrestling stunt on an elevated concert stage with a seated crowd in the background. The description is accurate in all key details: the number of participants, the color scheme of the stage, the presence of a crowd, and the nature of the stunt. **The answer is Yes.**  
 reject: The video does not accurately describe a wrestling stunt involving four people on a red-and-blue elevated concert stage with a seated crowd in front. **The answer is No.**

**Negative:**  
 desc: Four people do a wrestling stunt on a red-and-blue elevated concert stage with a **standing crowd** in front.  
 CoT: choose: The video shows four individuals participating in a wrestling stunt on a red-and-blue stage with an audience, but the description must be adjusted to account for the absence of a "standing crowd in front" detail. Therefore, the original statement is inaccurate. **The answer is No.**  
 reject: The description accurately captures the scene. There are indeed four people involved in a stunts, they are on a red-and-blue stage, and there appears to be a crowd watching from the foreground. Therefore, **the answer is Yes.**

**Temporal Reordering**

**Positive:**  
 desc: Four individuals on stage **first position themselves, then initiate the lift with two behind others.**  
 CoT: choose: The video shows a group of people, possibly wrestlers, lifting another individual from below to above their shoulders in a controlled manner, which aligns with the description provided. The text accurately describes the positioning and sequence of actions prior to the lift. **The answer is Yes.**  
 reject: The video depicts four individuals on stage, where one is lifted and flipped by another person, indicating a lift rather than a carry involving multiple people. The description in the question inaccurately states that two people lift someone over their shoulders, which does not align with the actual actions shown in the video. **The answer is No.**

**Negative:**  
 desc: Four individuals on stage **first execute the wrestling stunt, then initiate the lift with two behind others.**  
 CoT: choose: In the video, four individuals are seen on stage; however, only one lift is performed during the sequence. The video does not clearly depict that four individuals first execute a wrestling stunt and then the lift. **The answer is No.**  
 reject: The sequence of events in the video matches the description: four individuals execute a wrestling stunt with two then lifting one of them onto their shoulders. Therefore: **the answer is Yes.**



Bias Model



ORPO Training



Unbiased Model

Figure 25. Visualization of the hallucination preference data.

**Question**

What is the woman's intention of suddenly laughing at 0:40?

A: Conceal inner anger                      B: Think men's behavior is funny

C: Relieve the awkward atmosphere       D: Prepare to launch superpower punishment

E: Test men's reactions                      F: Express the joy of victory

**Video-RFT**

<think>

To determine the woman's intention when she suddenly laughs at 0:40, let's review the sequence of events in the video. Initially, the video shows a man sitting on a couch looking at his phone, while a woman stands behind him, also using her phone. The scene then shifts to the kitchen, where the man is seen drinking from a water bottle, appearing distressed.

Around 0:40, the woman is shown laughing, and the man is still in the kitchen, holding the water bottle. The video later reveals that **the man has been poisoned by the water he drank, leading to his collapse. The woman's laughter likely stems from her realization that she has tricked the man into drinking the poison.**

Given this context, the woman's laughter could be interpreted as a test of the man's reactions or a way to relieve the awkward situation. However, considering the options provided, Option C ("Relieve the awkward atmosphere") seems most fitting. The other options (A, B, D, E, F) do not align as well with the visual evidence. Therefore, based on the visual evidence, the woman's intention when she suddenly laughs at 0:40 is to relieve the awkward atmosphere. The answer is C.

</think>

✘ <answer>C</answer> Naive scene-centric analysis neglecting latent inter-event dependencies.

**VideoChat-R1.5**

<think>

At 0:40, the woman is seen laughing while holding a phone. **This could indicate that she finds something amusing on the phone, which might be related to the man's behavior or reaction.** Given the context of the video, where the man seems to be reacting to something the woman is showing him, it's likely that her laughter is a response to his actions or words. Therefore, the most fitting option would be that she thinks his behavior is funny.

</think>

✘ <answer>B</answer> Too brief, contains no reasoning or thought.

**Video-PLR**

<think>

To understand the woman's intention of suddenly laughing at 0:40, I need to identify the context leading up to this moment and observe her actions and interactions with others in the scene.

Initial Perception: The first relevant scene is <start="0.0s", end="6.0s", desc="A person is holding a smartphone with an Instagram post by a woman of attractive appearance displayed on the screen.">. This initial clip shows the man interacting with a smartphone. I still need to find evidence for what happened next, but it does not provide enough context.

Re-Perceive: To complete the reasoning, I will look at the next relevant scene. The next scene is <start="6.0s", end="30.0s", desc="A man is sitting on a sofa with his phone, interacting with a woman who approaches him. The woman sees something on his Instagram and takes the phone to look at it.">. This scene shows the woman is very curious about what's on the phone. However, it still doesn't explain why the woman would suddenly laugh. I need to see what happened next.

Re-Perceive: Now, I will examine the scene at <start="30.0s", end="68.0s", desc="The man and woman on the sofa are interacting, the woman laughs and touches the man's nose with her finger, their interaction still seems to be related to the phone.">. This scene continues to show the man and woman interacting around the phone. But the scene is gradually becoming tense. I need to find more clues.

Re-Perceive: Finally, the scene at <start="68.0s", end="157.0s", desc="The man becomes extremely distressed and eventually collapses to the ground, while the woman watches nearby.">. This scene shows the man's behavior is abnormal, and the woman does not offer any help, suggesting a malicious intent in the interaction.

Now I have all the necessary scenes to construct the sequence of events. **The woman's sudden laughter at 0:40 seems to be part of some kind of interaction pattern, and she has a strong sense of control over her environment. Combined with the content of the man's Instagram on his phone, the most plausible explanation is that the woman is preparing to unleash some kind of supernatural punishment on the man.**

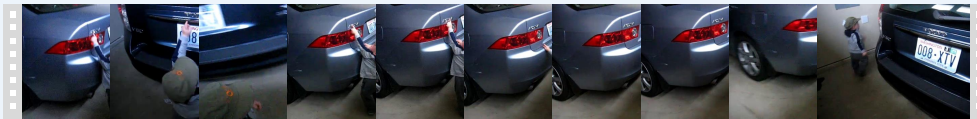
</think>

✔ <answer>D</answer> Continuous monitoring of video content and reasonable inference for implicit event capture.

Figure 26. **Case study on Video-Holmes.** Video-PLR leverages its perception loop to capture dispersed visual cues (e.g., the social media usage and the nose touch). This mechanism enables the model to reason about the **implicit causal relationships** between events, successfully identifying the woman's punitive intent where other models fail.



Video



0.0s 4.8s 9.6s 14.3s 19.1s 23.9s 28.7s 33.4s 38.2s 43.0s



How many cars appeared in the video?

A: 1

B: 2

Question C: 3

D: 4

#### Video-RFT

<think>

To determine how many cars appeared in the video, I need to recall the content of the video. The video shows a child interacting with a blue car in a dimly lit garage. Initially, the child is seen near the rear of the car, touching the license plate area. As the video progresses, the child continues to reach towards the back of the car, seemingly trying to attach or remove something from the license plate. No other vehicles are visible throughout the sequence.

The key elements in the video include the blue car, the child's actions, and the license plate. There are no indications of other cars being present. The focus remains solely on the interaction between the child and the single car. Therefore, based on the visual evidence, only one car is shown in the video. Given the options provided (A:1, B:2, C:3, D:4), the correct answer is Option A.

</think>

<answer>A</answer>



Non-iterative observation lacking granularity.

#### VideoChat-R1.5

<think>

There is only one car visible throughout the video. The child interacts with this single car, touching its rear end and license plate area. There are no other cars appearing in the video.

</think>

<answer>A</answer>



Non-iterative observation lacking granularity.

#### Video-PLR

<think>

To determine how many cars appeared in the video, I need to identify and count each car visible throughout the footage.

Initial Perception: The first relevant scene is <start="0.0s", end="3.0s", desc="A child interacts with the rear of a silver car in an indoor setting.">. This initial clip shows one car, but it only explains part of the story. I still need to find evidence for any additional cars that might appear later in the video.

Re-Perceive: To complete the reasoning, I will look at the next relevant scene. The next scene is <start="3.0s", end="10.0s", desc="The scene shifts to show another car, a dark-colored vehicle, in the background.">. This clip reveals a second car, which adds to the count. However, I still need to confirm if there are any more cars appearing in the remaining scenes.

Re-Perceive: Now, I will check the third relevant scene: <start="10.0s", end="43.0s", desc="The child continues to interact with the silver car, and finally the shot switches back to the dark-colored vehicle.">. This scene confirms the presence of the two cars that appeared earlier. It does not introduce any new cars.

Alright, I now have all the necessary clips to construct the sequence of events. Based on the evidence, two cars are visible in the video: the silver car and the dark-colored vehicle.

</think>

<answer>B</answer>



Multi-pass fine-grained inspection.

Figure 27. **Case study on VCR-Bench.** Although the video predominantly focuses on a single vehicle, a second car appears briefly (3s-10s). Compared to models that ignore these short segments, Video-PLR employs an iterative perception loop to conduct multi-pass fine-grained inspection, effectively mitigating critical information omission.

## Entity Name Extraction (From QA Text)

▶ **System Prompt:**

You are a helpful assistant.

▶ **User Prompt:**



**Task:**

Analyze the provided questions and answers to identify all specific visual elements, objects, or actions.



**Input:**

<Formatted QA Pairs>



**Constraints:**

- **DO NOT** output any reasoning or explanations.
- **DO NOT** output any temporal references (e.g., “at the beginning”, “near the end”).



**Guidelines:**

1. Your analysis must combine information from **BOTH** the question and the answer.
2. Extract **ONLY** tangible visual content (e.g., "man in red", "holding a balloon").



**Output Format:**

A concise list of the extracted visual elements.

Figure 28. **Step 1: Entity Name Extraction.** The prompt extracts visual entities from the original QA pairs to standardize terminology for subsequent captioning.

## Frame Level Captions (From Video Frames)

### ► System Prompt:

You are a helpful assistant.


### ► User Prompt:

#### 📋 Task:

Provide a detailed description of the objects that should be focused on. Output as **JSON** format.

#### 📄 Input:

<Input Frame  $F_t$ >  </Input Frame>

<Input Frame  $F_{t-1}$ >  </Input Frame>

Normalized Entity List <Normalized Entity List  $E$ >

Previous Caption <Caption  $C_{t-1}$ >

#### 📋 Guidelines:

##### 1. Overall

- Describe the frame and highlight changes from the first frame.
- Include details about moving or repositioned objects.

##### 2. Main Objects

###### ⚠️ Focus

- Only describe movable objects (e.g., people, animals, vehicles) in the frame.
- Include detailed descriptions of their appearance, actions, and associated objects.
- "Associated\_Objects" are items they interact with or relate to (e.g., "holding a cup", "standing beside a table").
- Clothing and accessories should be described under "appearance."

###### 👤 Appearance

- Provide concise but specific descriptions to avoid ambiguity.
- Examples: "A person wearing a yellow hazmat suit", "A child in a blue outfit."

###### 🔍 Representation

- Use "person", "man", or "woman" if most of the body is visible.
- For partial visibility, specify the body part (e.g., "hand holding a book", "head with a red hat").

###### ✅ Consistency

- Refer to object descriptions from the previous frame, but only include objects present in the frame.
- Ensure object names and relationships remain consistent across frames.

##### 3. Text And Stationary Objects

- Include all text objects (e.g., letters, logos) and stationary objects (e.g., buildings, trees).

#### 📄 Output JSON Format:

```
{
  "overall": "[Describe what happened in the frame]",
  "main_objects": [
    {
      "name": "[unique object name]",
      "associated_objects": [
        {
          "name": "[associated object name]",
          "relationship": "[relationship description]"
        }
      ]
    },
    {
      "name": "[unique object name]",
      "associated_objects": [
        {
          "name": "[associated object name]",
          "relationship": "[relationship description]"
        }
      ]
    }
  ],
  "appearance": "[concise appearance description]"
  "action": "[detailed action description]"
}
],
"text_and_stationary_objects": "[text_and_stationary_objects in the frame]"
}
```

Figure 29. **Step 2: Frame Level Captioning.** The prompt generates detailed descriptions for the current frame while maintaining object consistency with the previous frame.

### Initial CoT Generation (rom Structured Text)

#### ► System Prompt:

You are a video analyst capable of breaking down videos into multiple events, with each event having approximately the same duration.

#### ► User Prompt:

##### 📝 Task:

Your task is to generate a step-by-step reasoning process (Chain of Thought) that logically connects synthesized "events" from the video. Your response must **ONLY** be this reasoning process. Do not include any other text, explanations, or markdown formatting.

##### 📥 Input:

- A string containing descriptions of video frames, where each moment is enclosed in <moment> tags: <videos>. (from step2 Frame Level Captions)
- A question about the video: <question>.
- The correct answer to the question: <answer>.

##### 📋 Core Instructions:

Your reasoning process must follow this exact iterative structure:

1. **Plan:** Begin by analyzing the <question> **ONLY**. State what sequence of events you need to find in the video to logically solve the problem.
2. **Iterative Evidence Gathering:**
  - Enter a dynamic loop of examining evidence and assessing its sufficiency.
  - **Perception:** Present a key synthesized event. This event must be derived from a single frame or by summarizing a change across multiple consecutive frames. You **MUST** use the format: <start= "x.x s", end= "x.x s", desc= "A concise description of the key event.">
  - **Assess Evidence:** Immediately after presenting the event, assess it. Explain what this event tells you, and whether it's sufficient. State what is still missing and what you need to look for next.
  - **Repeat:** Continue this Perception/Assess loop (2-4 times) until you have gathered enough evidence to fully support the answer.
3. **Summary and Analysis:** Once you determine the evidence is sufficient, provide a comprehensive analysis. Gather all the events you found, explain the chronological and logical connections between them, and show how this chain of events definitively leads to the provided answer.

##### ⚠️ Constraints:

- **Grounding:** Your entire reasoning must be based **EXCLUSIVELY** on the textual information provided in the video descriptions. Do not invent visual details.
- **Event Synthesis:** An "event" synthesizes information. It can be derived from a single frame description or by summarizing a change across multiple consecutive frame descriptions.
- **Output:** Your entire response must be **ONLY** the reasoning process outlined above, starting with 'CoT:' and ending with 'Answer:'.
- **Reasoning without knowing the answer:** It is necessary to start reasoning from the unknown, and during the reasoning process, it is not allowed to imply or explicitly rely on the wrong logic of the standard answer.

##### 📌 Note

1. The evidence for each event needs to last for at least 6 seconds, that is, start\_time - end\_time > 6.
2. Keep the language concise. Don't over-think. The number of events is no more than 3.

##### 📄 Output Format:

CoT: [Your detailed reasoning process following the structure above]

Answer: [The provided correct answer]

##### Example:

###### Input Data (Simplified):

Descriptions of video frames: xxx video desc (from step2 Frame Level Captions)

Question: Why did the rabbit run?

Answer: It saw a predator from above.

###### Output:

**CoT:** To solve this problem, I need to observe the actions of the rabbit and its reactions before and after running.

**Perception:** <start= "0.0s", end= "6.5s", desc= "The video descriptions show a brown rabbit calmly "chewing on grass in a field," which establishes its baseline behavior.">

**Assess Evidence:** This establishes the rabbit's normal, calm state. This is not enough to explain why it ran. I need to find the stimulus event that changed this behavior.

**Perception:** <start= "4.1s", end= "15.1s", desc= "The descriptions state a 'large shadow moves over the rabbit'. Following this, the rabbit's 'ears are up,' and it is 'looking up,' indicating it has perceived a threat from above.">

**Assess Evidence:** This is the key stimulus. The "large shadow" combined with the reaction of "looking up" directly supports the "from above". This is the cause. I now need to find the final effect.

**Perception:** <start= "14.0s", end= "26.5s", desc="The descriptions confirm the rabbit's action is now 'running quickly across the field.'">

**Assess Evidence:** This is the final piece of the puzzle. After perceiving the shadow and looking up, the rabbit's action is "running quickly." This completes the logical chain. I have sufficient evidence.

**Summary and Analysis:** The synthesized events, when combined, provide a clear causal link.

The chronological order of the events (calm baseline state → shadow above → running) logically suffices to show that the answer is "rabbit saw a predator from above".

**Answer:** It saw a predator from above.

Figure 30. **Step 3: Initial CoT Generation.** The prompt synthesizes frame-level descriptions into a structured reasoning chain with timestamps (Start, End, Description) and sufficiency assessment.

## CoT Logic Refinement (Text-to-Text)

### ► System Prompt:

You are a helpful assistant.

### ► User Prompt:



#### Task:

You are tasked with rewriting the "think" section of given data to enhance the chain of thought (CoT) process. Rewrite the "think" content based on the given "question" and original "think" to meet these requirements.



#### Input:

- Input Question: <question>
- Origin Think Section: <think> (from step3 Initial CoT Generation)



#### Guidelines:

**1. Initial Analysis:** Analyze the problem and formulate what visual information should be observed in the video in order to solve the problem. Avoid revealing the answer here.

**2. Perception & Assess Evidence:** For each perception segment with <start= "x.x s", end= "x.x s", desc= description>:

- After analyzing the description, explicitly evaluate possible options with a hypothesis. Use a pattern like: "I would consider [Option X] as a possible answer, but I need to check other parts of the video for more evidence" or "Based on this, [Option X] seems plausible, but further verification is needed".
- If the current information is insufficient to confirm any option, state that and explain what more is needed.

**3. Summary and Analysis:**

- Synthesize all perceptions to verify the correct answer and explicitly exclude other options.
- For the correct option: Explain how all evidence supports it.
- For excluded options: Clearly state why they are invalid (e.g., no evidence supports them, or evidence contradicts them).
- Maintain a logical flow showing how the evidence leads to the conclusion.

**4. Format Preservation**

- Keep the exact structure: Start with "Cot:", followed by the structured reasoning (Perception → Assess Evidence → ... → Summary and Analysis), and end with "Answer: [Final Answer]".
- Do not modify the original video perception timestamps and descriptions: <start= "x.x s", end= "x.x s", desc=...>.
- Ensure the revised reasoning is coherent, analytical, and directly addresses the question.



#### Output Format:

**CoT:** [A refined CoT data from step3 Initial CoT Generation]

**Answer:** [The provided correct answer]

Figure 31. **Step 4: CoT Logic Refinement.** The prompt refines the initial chain of thought to ensure the logical flow is sound and explicitly excludes incorrect options based on visual evidence.

### Detailed Factual Recaptioning (Video-to-Text)

► **System Prompt:**

You are a helpful assistant.

► **User Prompt:**

📄 **Task:**

Based on the central caption of  $\langle \text{event} \rangle$ , provide a concise and factual description of the video.

📄 **Input:**

$\langle \text{Input Frame } F_{start\_time} - F_{end\_time} \rangle$  📺  $\langle \text{Input Frame} \rangle$

📖 **Guidelines:**

Focus on these specific, observable details:

1. **People's Features:** Briefly note the key visual characteristics of the people involved (e.g., clothing, hair).
2. **Action Sequence:** Describe the main actions in the order they occur.
3. **Objects:** Mention any significant objects, describing their appearance and how they are used.

⚠️ **Note:**

Stick strictly to what you see. Do not describe the overall setting or make assumptions.

Figure 32. **Detailed Factual Recaptioning.** The prompt used to generate high-fidelity dense descriptions from video clips, serving as the ground truth ( $C_{gt}$ ) for the hallucination dataset.

## Initial CoT Generation (From Structured Text)

### ► System Prompt:

You are a video analyst capable of breaking down videos into multiple events, with each event having approximately the same duration.

### ► User Prompt:

#### 📄 Task:

Your task is to generate a step-by-step reasoning process (Chain of Thought) that logically connects synthesized "events" from the video. Your response must **ONLY** be this reasoning process. Do not include any other text, explanations, or markdown formatting.

#### 📄 Input:

- A string containing descriptions of video frames, where each moment is enclosed in <moment> tags: <videos>. (from step2 Frame Level Captions)
- A question about the video: <question>.
- The correct answer to the question: <answer>.

#### 📄 Core Instructions:

Your reasoning process must follow this exact iterative structure:

1. **Plan:** Begin by analyzing the <question> **ONLY**. State what sequence of events you need to find in the video to logically solve the problem.
2. **Iterative Evidence Gathering:**  
Enter a dynamic loop of examining evidence and assessing its sufficiency.
  - **Perception:** Present a key synthesized event. This event must be derived from a single frame or by summarizing a change across multiple consecutive frames. You **MUST** use the format: <start= "x.x s", end= "x.x s", desc= "A concise description of the key event.">
  - **Assess Evidence:** Immediately after presenting the event, assess it. Explain what this event tells you, and whether it's sufficient. State what is still missing and what you need to look for next.
  - **Repeat:** Continue this Perception/Assess loop (2-4 times) until you have gathered enough evidence to fully support the answer.
3. **Summary and Analysis:** Once you determine the evidence is sufficient, provide a comprehensive analysis. Gather all the events you found, explain the chronological and logical connections between them, and show how this chain of events definitively leads to the provided answer.

#### ⚠️ Constraints:

- **Grounding:** Your entire reasoning must be based **EXCLUSIVELY** on the textual information provided in the video descriptions. Do not invent visual details.
- **Event Synthesis:** An "event" synthesizes information. It can be derived from a single frame description or by summarizing a change across multiple consecutive frame descriptions.
- **Output:** Your entire response must be **ONLY** the reasoning process outlined above, starting with 'CoT:' and ending with 'Answer:'.
- **Reasoning without knowing the answer:** It is necessary to start reasoning from the unknown, and during the reasoning process, it is not allowed to imply or explicitly rely on the wrong logic of the standard answer.

#### 📄 Note

1. The evidence for each event needs to last for at least 6 seconds, that is, start\_time - end\_time > 6.
2. Keep the language concise. Don't over-think. The number of events is no more than 3.

#### 📄 Output Format:

CoT: [Your detailed reasoning process following the structure above]

Answer: [The provided correct answer]

#### Example:

##### Input Data (Simplified):

Descriptions of video frames: xxx video desc (from step2 Frame Level Captions)

Question: Why did the rabbit run?

Answer: It saw a predator from above.

##### Output:

**CoT:** To solve this problem, I need to observe the actions of the rabbit and its reactions before and after running.

**Perception:** <start= "0.0s", end= "6.5s", desc= "The video descriptions show a brown rabbit calmly 'chewing on grass in a field', which establishes its baseline behavior.">

**Assess Evidence:** This establishes the rabbit's normal, calm state. This is not enough to explain why it ran. I need to find the stimulus event that changed this behavior.

**Perception:** <start= "4.1s", end= "15.1s", desc= "The descriptions state a 'large shadow moves over the rabbit'. Following this, the rabbit's 'ears are up', and it is 'looking up', indicating it has perceived a threat from above.">

**Assess Evidence:** This is the key stimulus. The "large shadow" combined with the reaction of "looking up" directly supports the "from above". This is the cause. I now need to find the final effect.

**Perception:** <start= "14.0", end= "26.5s", desc= "The descriptions confirm the rabbit's action is now 'running quickly across the field.'">

**Assess Evidence:** This is the final piece of the puzzle. After perceiving the shadow and looking up, the rabbit's action is "running quickly." This completes the logical chain. I have sufficient evidence.

**Summary and Analysis:** The synthesized events, when combined, provide a clear causal link.

The chronological order of the events (calm baseline state -> shadow above -> running) logically suffices to show that the answer is "rabbit saw a predator from above".

**Answer:** It saw a predator from above.

Figure 33. **Negative Caption Generation.** The prompt instructs the model to generate a specific type of hallucinated caption (e.g., Attribute Modification) based on the ground truth.

## Frame Level Captions (From Video Frames)

### ► System Prompt:

You are a helpful assistant.

### ► User Prompt:

#### 🗒️ Task:

Provide a detailed description of the objects that should be focused on. Output as **JSON** format.

#### ➡️ Input:

<Input Frame  $F_t$ > 🖼️ </Input Frame>

<Input Frame  $F_{t-1}$ > 🖼️ </Input Frame>

Normalized Entity List <Normalized Entity List  $E$ >

Previous Caption <Caption  $C_{t-1}$ >

#### 📖 Guidelines:

##### 1. Overall

- Describe the frame and highlight changes from the first frame.
- Include details about moving or repositioned objects.

##### 2. Main Objects

###### ⚠️ Focus

- Only describe movable objects (e.g., people, animals, vehicles) in the frame.
- Include detailed descriptions of their appearance, actions, and associated objects.
- "Associated\_Objects" are items they interact with or relate to (e.g., "holding a cup", "standing beside a table").
- Clothing and accessories should be described under "appearance."

###### 👤 Appearance

- Provide concise but specific descriptions to avoid ambiguity.
- Examples: "A person wearing a yellow hazmat suit", "A child in a blue outfit."

###### 🔍 Representation

- Use "person", "man", or "woman" if most of the body is visible.
- For partial visibility, specify the body part (e.g., "hand holding a book", "head with a red hat").

###### ✅ Consistency

- Refer to object descriptions from the previous frame, but only include objects present in the frame.
- Ensure object names and relationships remain consistent across frames.

##### 3. Text And Stationary Objects

- Include all text objects (e.g., letters, logos) and stationary objects (e.g., buildings, trees).

#### 📄 Output JSON Format:

```
{
  "overall": "[Describe what happened in the frame]",
  "main_objects": [
    {
      "name": "[unique object name]",
      "associated_objects": [
        {
          "name": "[associated object name]",
          "relationship": "[relationship description]"
        }
      ],
      "appearance": "[concise appearance description]"
      "action": "[detailed action description]"
    }
  ],
  "text_and_stationary_objects": "[text_and_stationary_objects in the frame]"
}
```

Figure 34. **Positive Caption Generation (Correction).** The prompt asks the model to correct a negative caption back into a factual description, ensuring paired sentence structures.

## CoT Logic Refinement (Text-to-Text)

### ► System Prompt:

You are a helpful assistant.

### ► User Prompt:

#### Task:

You are tasked with rewriting the "think" section of given data to enhance the chain of thought (CoT) process. Rewrite the "think" content based on the given "question" and original "think" to meet these requirements.

#### Input:

- Input Question: <question>
- Origin Think Section: <think> (from step3 Initial CoT Generation)

#### Guidelines:

**1. Initial Analysis:** Analyze the problem and formulate what visual information should be observed in the video in order to solve the problem. Avoid revealing the answer here.

**2. Perception & Assess Evidence:** For each perception segment with <start= "x.x s", end= "x.x s", desc= "A concise description of the key event.">

- After analyzing the description, explicitly evaluate possible options with a hypothesis. Use a pattern like: "I would consider [Option X] as a possible answer, but I need to check other parts of the video for more evidence" or "Based on this, [Option X] seems plausible, but further verification is needed".
- If the current information is insufficient to confirm any option, state that and explain what more is needed.

**3. Summary and Analysis:**


- Synthesize all perceptions to verify the correct answer and explicitly exclude other options.
- For the correct option: Explain how all evidence supports it.
- For excluded options: Clearly state why they are invalid (e.g., no evidence supports them, or evidence contradicts them).
- Maintain a logical flow showing how the evidence leads to the conclusion.

**4. Format Preservation**

- Keep the exact structure: Start with "CoT:", followed by the structured reasoning (Perception → Assess Evidence → ... → Summary and Analysis), and end with "Answer: [Final Answer]".

- Do not modify the original video perception timestamps and descriptions:

<start= "x.x s", end= "x.x s", desc= "...">

-  Ensure the revised reasoning is coherent, analytical, and directly addresses the question.

#### **Output Format:**

**CoT:** [A refined CoT data from step3 Initial CoT Generation]

**Answer:** [The provided correct answer]

Figure 35. **CoT Judgement Prompt.** The prompt used for the Factual-Aware Evaluator (FAE) to determine whether a text description matches the video using a reasoning step.

### PLR Think Prompt

► **PLR Think Prompt:**

The assistant first thinks about the briefly reasoning process with evidence and then provides the user with the answer.

Within the `<think></think>` tags, The assistant provide reasoning with visual evidence: `<start= "x.x s", end= "x.x s", desc= "briefly explains the evidence\">`.

The assistant should gradually find visual evidence and evaluate whether it is sufficient to answer the question, and end the thinking after multiple loops. Then, provide the answer within the `<answer></answer>` tags.

Figure 36. **Video-PLR Thinking Prompt.** The inference prompt that instructs the model to use the `<think>` tag and perform the perception loop mechanism before answering.