When Thinking Drifts: Evidential Grounding for Robust Video Reasoning

Mi Luo[1] **Zihui Xue**[1] **Alex Dimakis**[2, 3] **Kristen Grauman**[1]
[1]The University of Texas at Austin [2]UC Berkeley [3]Bespoke Labs

Abstract

Video reasoning, the task of enabling machines to infer from dynamic visual content through multi-step logic, is crucial for advanced AI. While the Chain-of-Thought (CoT) mechanism has enhanced reasoning in text-based tasks, its application to video understanding remains underexplored. This paper presents a systematic analysis revealing that CoT often degrades performance in video reasoning, generating verbose but misleading internal monologues, and leading to hallucinated visual details and overridden correct intuitions—a phenomenon we term "visual thinking drift." We explain this drift through a Bayesian lens, positing that CoT traces often diverge from actual visual evidence, instead amplifying internal biases or language priors, causing models to storytell rather than engage in grounded reasoning. To counteract this, we introduce Visual Evidence Reward (VER), a reinforcement learning framework that explicitly rewards the generation of reasoning traces that are verifiably grounded in visual evidence. Comprehensive evaluation across 10 diverse video understanding benchmarks demonstrates that our Video-VER consistently achieves top performance. Our work sheds light on the distinct challenges of video-centric reasoning and encourages the development of AI that robustly grounds its inferences in visual evidence—for large multimodal models that not only "think before answering", but also "see while thinking".[1]

1 Introduction

Imagine watching a cooking tutorial video: as the chef chops vegetables, combines ingredients, and adjusts the heat, we are not just passively observing isolated actions. Our minds actively connect these steps, anticipating the next move, understanding the purpose behind each technique, and picturing the final dish. This act of deriving understanding from a sequence of visual events, inferring intent, and predicting results through a series of logical thoughts mirrors the core challenge addressed by *video reasoning*: empowering machines to draw inferences and conclusions from dynamic content of video through multi-step logical reasoning [16, 41, 43, 17, 55, 42].

As researchers strive to imbue systems with this capability, their efforts have largely branched into two complementary strategies. One prominent path is the agentic workflow, also referred to as modular reasoning [42], which deconstructs video understanding into focused sub-problems—such as event parsing, object tracking, and temporal grounding addressed by specialized models [68, 15, 55, 37, 42]. Alongside this modular approach, a parallel evolution in Multimodal Large Language Models (MLLMs) has ushered in a more holistic reasoning paradigm, where a model executes end-to-end inference by harnessing both vision and text [5, 29, 75, 69, 72, 10, 34, 31, 9]. Within this architecture, the vision encoder typically assumes the role of a perceptual engine, tasked with extracting high-fidelity visual representations, whereas the language model component takes the helm of the reasoning process itself, generating textual rationales and constructing logical chains.

[1]Project webpage: https://vision.cs.utexas.edu/projects/video-ver/.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

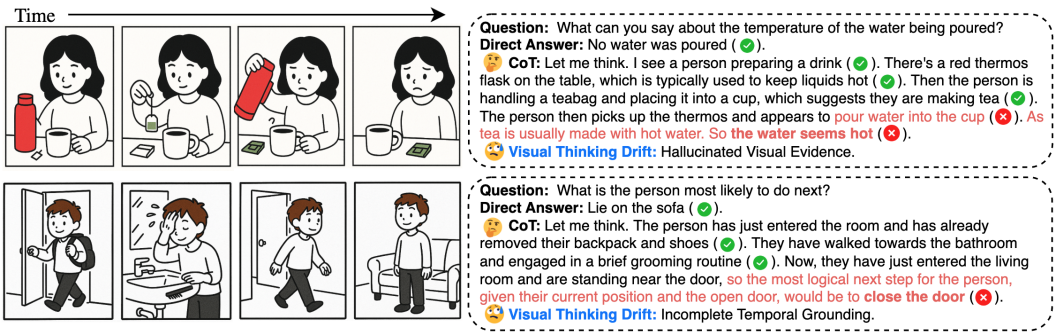


Figure 1: Two examples of *Visual Thinking Drift* phenomenon, where the reasoning chain, as it grows longer, increasingly relies on hallucinated facts or incomplete temporal context—drawing conclusions from language patterns rather than grounding in the actual video content.

Building upon this foundational capability, a significant body of recent work explores chain-of-thought (CoT) reasoning for MLLMs, both through high-quality reasoning datasets with spatio-temporal annotations [48, 39, 22] as well as reinforcement learning (RL) post-training approaches inspired by influential image- and text-based reasoners [21, 23, 66, 51, 58] that refine the MLLM's reasoning pathways [17, 74, 33] with rule-based rewards. Early results suggest that simply encouraging "thinking before answering" can often pay off. However, despite these promising advancements, the unique challenges encountered when transposing *text*-based chain-of-thought reasoning to the distinct demands of *video*-centric tasks warrant deeper exploration.

In this paper, we aim to both expose the gaps in CoT-based video reasoning, and propose a solution. First we present a systematic study covering 10 video benchmarks, multiple MLLMs, and 20 video QA subtasks, revealing that CoT reasoning often backfires for video understanding, especially with open-sourced models. We identify a recurring failure mode we term "*Visual Thinking Drift*" where an MLLM introduces hallucinated facts or bias to outdated frames for temporal reasoning. Rather than enhance reasoning, the CoT prompts frequently induce models to produce verbose but misleading internal monologues—hallucinating visual details, overriding correct instincts, and ultimately reducing accuracy. See Figure 1. For instance, in next-action prediction tasks, the model may base its logic on earlier events while ignoring more recent cues—despite being able to answer correctly when prompted directly, without CoT. This drift reveals a critical flaw: CoT reasoning in video models often becomes performative rather than grounded—fluent, plausible, but ultimately wrong. To understand this phenomenon, we adopt a Bayesian lens, showing that CoT traces often diverge from actual visual evidence and instead amplify internal biases or language priors.

Next, to counter the visual thinking drift problem, we introduce Visual Evidence Reward (VER), a novel reward mechanism for reinforcement learning-based MLLM post-training framework. Our VER explicitly encourages reasoning traces grounded in visual evidence. The key insight is that genuine video reasoning emerges when the internal thought process itself is actively and granularly tethered to perceived content, compelling models to truly "see while thinking", not just "think before answering". In the proposed model, an auxiliary LLM acts as a judge, evaluating the factual alignment between intermediate thoughts and visual inputs. This automatically encourages not just coherent but correct reasoning, stabilizing inference, and boosting overall accuracy.

We evaluate our Video-VER model across 10 diverse video understanding benchmarks. Compared to strong base models and existing reasoning techniques, Video-VER consistently ranks first or second. Furthermore, our model achieves consistently strong margins compared to its respective base MLLM (trained without the Visual Evidence Reward)—as much as +9.0% absolute accuracy gains, and an average of +4.0% across all 10 benchmarks. Our results suggest that in video reasoning, grounding—not verbosity—is essential to true video intelligence.

2 Related Work

Eliciting Reasoning Ability from Large Language Models Large language models (LLMs) have shown strong performance on complex reasoning tasks such as mathematics and programming [7, 11, 50, 2, 47]. These capabilities are often elicited through few-shot [58, 8, 64, 78] and zero-shot prompting [65, 27], or through instruction tuning with large-scale chain-of-thought (CoT)

datasets [12, 44, 52]. Recent advances show that even simple rule-based incentive mechanisms and lightweight reinforcement learning can induce robust reasoning without explicit supervision [21]. However, studies also highlight that LLM-generated reasoning traces may be unreliable or unfaithful to the model's internal processes [46, 53]. Motivated by these insights, we investigate how to elicit reasoning abilities from multimodal LLMs for video understanding—a domain that introduces unique challenges due to the dynamic temporal nature of video data.

Video Reasoning Video reasoning entails drawing conclusions from video content through multi-step logical inference [16]. As overviewed above, research in this area follows two directions: modular reasoning [42] that decomposes tasks into addressable subcomponents [68, 15, 55, 37, 42] and MLLMs that perform end-to-end inference by jointly leveraging visual and textual information [5, 29, 75, 69, 72, 10, 34, 31, 9]. Building on this, some recent work explores enhanced chain-of-thought (CoT) reasoning for MLLMs. One line focuses on constructing high-quality reasoning datasets, grounded temporally or spatially, to guide more structured logic generation [48, 39, 22]. Another emerging direction, inspired by DeepSeek-R1 [21], Open Reasoner Zero [23], Skywork R1V [66] and Kimi k1.5 [51], applies reinforcement learning (RL) [49, 40] to refine the reasoning process through lightweight, targeted reward mechanisms [17, 74, 33]. Despite these promising developments, systematic analyses of challenges of text-based chain-of-thought reasoning in video tasks remain limited. We contribute to this space by offering both empirical insights and a simple yet effective reward strategy designed to improve the faithfulness to visual content for reasoning chains.

Hallucination in MLLMs We identify a novel phenomenon termed "*Visual Thinking Drift*", a specific manifestation of hallucination in MLLMs. While hallucination—producing descriptions, or conclusions misaligned with visual input—has long been a challenge for MLLMs [6], visual thinking drift is distinguished by its emergence within chain-of-thought reasoning: errors introduced at earlier reasoning steps, once hallucinated, can propagate through subsequent steps, ultimately leading to conclusions that significantly diverge from the visual evidence. In the image domain, prior research primarily focuses on object hallucination, where models misidentify object categories, attributes, or relationships [6]. In the video domain, hallucinations involve misinterpretations of dynamic actions, events, and narrative sequences [57, 70]. To mitigate such issues, existing work explores test-time interventions [36, 54, 25, 30], preference modeling during training [60, 73, 62] and progress-aware data filtering [61] to reduce vision-language misalignment. In contrast, we propose a lightweight and data-efficient alternative: reinforcement fine-tuning to mitigate hallucination within the reasoning process itself. Rather than focus solely on perception-level correction, our approach targets the integrity of logical progression, aiming to curb the cascading effects of visual thinking drift.

3 Dilemma of Chain-of-Thought Reasoning in Video Understanding

The standard approach to evaluating Video LLMs (a.k.a., MLLMs) for Video Question Answering (VQA) [32, 18] focuses on their ability to provide direct answers—such as returning an integer for a question like "How many objects enter the scene?" Recent advances in LLM reasoning, particularly Chain-of-Thought (CoT) prompting [58], have encouraged models to reason step by step. This approach offers benefits in decomposing complex questions and improving the explainability of responses. In this study, we aim to systematically evaluate the two core prompting strategies: *Direct Response Generation* and *Reasoning-Driven Generation (Chain-of-Thought)*.

Direct Response Generation Under this scheme, the video LLM produces the output answer a by directly leveraging the input question q and the accompanying video context $\mathbf{v}$. The generation task can be formally expressed as:

$$p(a \mid q, \mathbf{v}).$$

In this approach, the model is expected to yield the final answer immediately, without constructing any intermediate reasoning path or justification.

Reasoning-Driven Generation (Chain-of-Thought) Conversely, this method decomposes the output generation into two sequential phases. Initially, the model infers a rationale sequence $c_{1:T}$ based on the input query q and the contextual video data $\mathbf{v}$. Subsequently, it conditions the final prediction a on both the generated rationale and the original inputs. This approach is captured by the following formulation:

$$p(c_{1:T} \mid q, \mathbf{v}) \cdot p(a \mid c_{1:T}, q, \mathbf{v}).$$

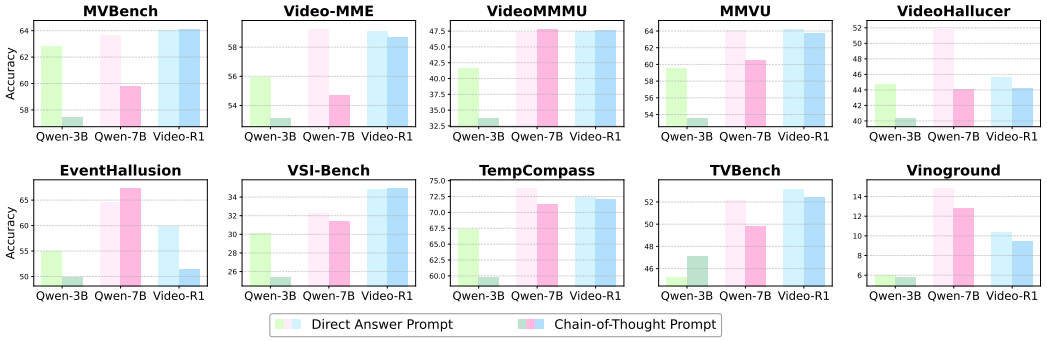


Figure 2: Compared to directly prompting the model for an answer, instructing the model to "think before answering" leads to a noticeable performance drop in open-source MLLMs such as Qwen2.5-VL [5] and Video-R1 [17] across multiple benchmarks (10 are shown here).

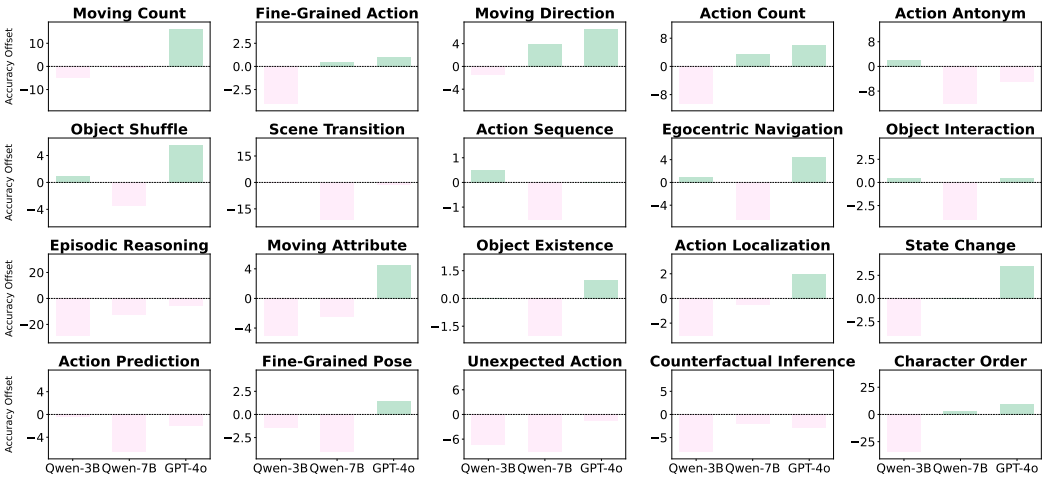


Figure 3: Gains (green) and losses (pink) with CoT prompt, showing that reasoning-driven generation is valuable for multi-hop, causal, or interpretability-driven tasks like object counting, but weakens both large and small models on lightweight perceptual questions such as scene transition detection.

3.1 How Does CoT Make Models Weaker on Simple Video Perception Tasks?

To assess the impact of reasoning-driven generation on video understanding, we conduct a systematic study comparing it to direct response generation. Our key question: Does prompting models to "think step-by-step" improve performance in state-of-the-art Video LLMs?

As illustrated in Figure 2, we evaluate both generation strategies across three leading open-sourced MLLMs—Qwen2.5-VL-3B [5], Qwen2.5-VL-7B [5], and Video-R1-7B [17]—on 10 diverse video benchmarks, spanning general video understanding [18] to complex temporal [71] and spatial reasoning tasks [63].[2] Surprisingly, reasoning-driven (CoT) prompting often leads to lower accuracy, particularly on benchmarks demanding rapid visual perception, like Video-MME [18].

To further dissect the impact of CoT prompting, we analyze 20 subtasks from MVBench (Figure 3), leveraging its structured task taxonomy. The results show that forcing models to "think out loud" *reduces* accuracy on tasks that rely on quick visual yes/no or single-label judgments, such as scene transition detection. The additional reasoning tokens can sometimes encourage over-rationalization, introduce hallucinated details, or dilute contextual focus—turning what should be a straightforward perceptual judgment into an unnecessarily deliberative process. Conversely, CoT reasoning can improve accuracy on tasks that genuinely benefit from structured decomposition, such as those involving multi-step spatial or temporal reasoning (e.g., counting objects and actions). As shown in Figure 4, even with the much larger proprietary model GPT-4o—a strong model with advanced reasoning capabilities—a significant portion of questions are better answered directly than with CoT.

[2]We also experimented with LLaVa-OneVision-7B [29], but it failed to follow the instruction to generate a thought trace, likely due to its use of a weaker underlying language model.

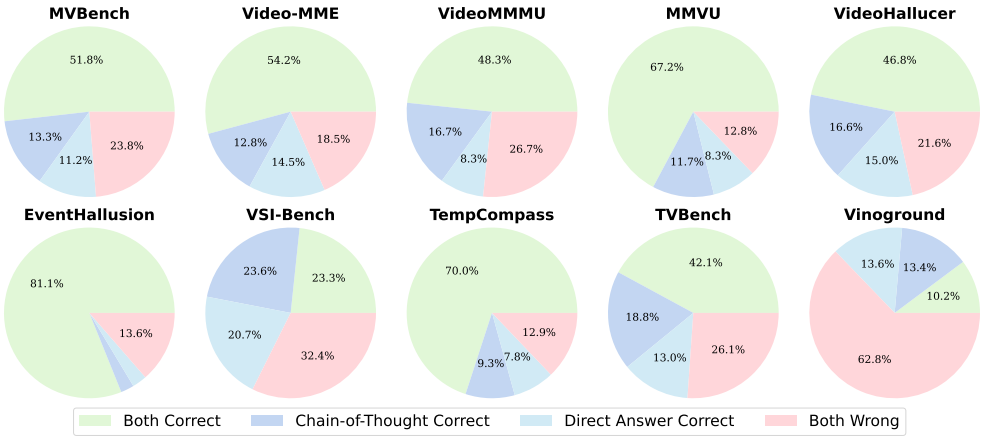


Figure 4: Even with GPT-4o (a strong reasoning model), a considerable portion of questions (light blue areas) are better answered directly than with CoT reasoning, implying significant room for improvement in CoT reasoning. For VSI-Bench and MMVU, results are based on MCQ subset.

In summary, while reasoning-driven generation tends to help tasks that benefit from explicit decomposition, its effects are not uniformly positive—it can hinder performance on simpler perceptual tasks and even some multi-step reasoning tasks.

3.2 Visual Thinking Drift: When Reasoning Ignores the Video

Reasoning implicitly unfolds in two stages: first, identifying the relevant rules and facts needed to reach a conclusion, and second, applying them effectively to arrive at that conclusion [28]. Simply encoding knowledge is not enough—robust reasoning under uncertainty is essential. However, the verbose nature of CoT traces introduces stochasticity into the reasoning process. Drawing inspiration from self-consistency [56], we found that majority voting over 20 responses generated with CoT prompt significantly improves accuracy (see supplementary material for details). Yet, this improved stability comes at the cost of considerable computational overhead.

To investigate the source of this instability, we analyzed multiple erroneous chains of thought. Paradoxically, we found that many flawed thinking traces in video analysis are logically flawless. The culprit? A phenomenon we call "**Visual Thinking Drift**", illustrated in Figure 1. The model's reasoning is sound, but it is unmoored from the video's true content—building its logic on hallucinated visual details or isolated temporal fragments, which inevitably steer the inference off track.

To better understand the Visual Thinking Drift phenomenon, we adopt a Bayesian lens, which helps disentangle why CoT can damage a video LLM's answer even when the direct answer alone is correct. Consider a video LLM that, given a question q and video features $\mathbf{v}$, generates a chain of reasoning tokens $c_{1:T}$ followed by a final answer a. Its implicit generative story is

$$p(c_{1:T}, a \mid q, \mathbf{v}) \;=\; p\big(a \mid c_{1:T}, q, \mathbf{v}\big) \prod_{t=1}^{T} p\big(c_t \mid c_{<t}, q, \mathbf{v}\big).$$

Because the chain tokens are never supervised, each inference step samples a high-variance *latent* state. Early in the generation process, visual evidence does influence the softmax

$$p\big(c_t \mid c_{<t}, q, \mathbf{v}\big) \;\propto\; \exp\big(\underbrace{\mathbf{h}_{c_{<t}}^{\top} W_{\text{lang}}}_{\text{language prior}} + \underbrace{\mathbf{h}_{\mathbf{v}}^{\top} W_{\text{vis}}}_{\text{visual likelihood}}\big),$$

yet in today's video-LLMs, the language-projection matrix tends to have a much larger norm than its visual counterpart — $\|W_{\text{lang}}\| \gg \|W_{\text{vis}}\|$. This is because the model inherits a massive, pre-trained language backbone trained on trillions of text tokens, while the vision pathway is typically a small, newly added adapter, fine-tuned on multimodal data that is orders of magnitude smaller. Moreover, during inference, far fewer visual tokens influence gradient flow compared to language tokens. This asymmetry in initialization, data scale, and gradient dynamics leads to a significantly smaller Frobenius norm in W_{vis}, causing language priors to dominate during autoregressive decoding and contributing to the observed drift. As t grows, self-attention focuses ever more on the already-generated tokens,

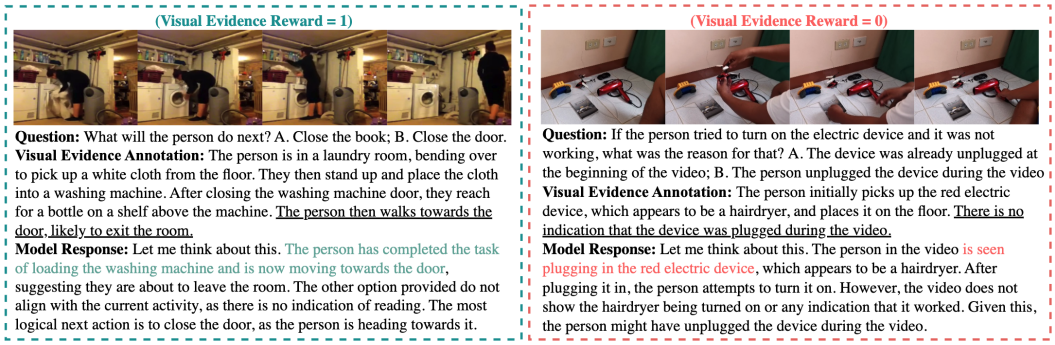


Figure 5: Visualization of visual facts generated from the training data. Chain-of-thought responses that actively reference visual evidence are rewarded, while those that do not receive zero reward.

so the visual likelihood is *diluted*. A linguistically plausible but ungrounded narrative emerges, and a single hallucinated detail can dominate all subsequent probabilities. If each individual step is correct with probability $p_s = 1 - \varepsilon$, the probability that an entire chain of length T is error-free is $(1-\varepsilon)^T \approx 1 - T\varepsilon$ for small ε; thus the failure rate grows *linearly* with chain length. Once an early token commits to a nonexistent visual fact (e.g. "the man holds a green ball"), all later tokens and the answer are conditioned on this fiction. Because autoregressive decoding has no backward message-passing to re-verify the video, the posterior mass collapses around the hallucination and recovery becomes nearly impossible.

Crucially, the model's high-probability *logical scaffolds*—"if–then" structures, counting loops, temporal ordering—stay intact, while the low-entropy visual details are weakly weighted. The result is a chain of thought that *looks* logically sound but is built on visually hallucinated content: the essence of *visual thinking drift*. For a detailed theoretical explanation of the Bayesian mechanism, please refer to the Appendix.

We show two concept examples of visual thinking drift made by CoT prompting in Figure 1, where CoT brings hallucinations or mismatches with visual evidence. When inconsistencies arise, MLLMs faithfully trust textual data over visual data, leading to wrong reasoning paths.

4 Visual Evidence Reward (VER) for MLLM Video Reasoning

Continuing this Bayesian perspective, a key insight into the visual thinking drift dilemma is that CoT tokens are never explicitly supervised during training. Without clear guidance, reasoning can easily drift away from the visual evidence it is meant to be grounded in. To tackle this, we enhance the lightweight rule-based reinforcement learning algorithm, Group Relative Policy Optimization (GRPO) [47, 21], by introducing a novel reward mechanism: Visual Evidence Reward (VER). VER directly supervises the model's reasoning process by rewarding it when its chain-of-thought includes accurate and relevant visual details—effectively anchoring abstract reasoning in concrete visual facts.

For each question q, we have a policy model π_θ to generate a group of responses $\{o_i\}_{i=1}^G$, where G is the group size. A large language model (LLM)-based judge evaluates each response o_i for its reference to the visual evidence $\mathbf{v}$, assigning a binary score $e_i \in \{0, 1\}$, where 1 indicates a proper reference. We define the evidence reward coefficient as $r_e = \alpha$ if $e_i = 1$, and $r_e = 0$ otherwise, where α is a tunable reward weight. Note that using an auxiliary LLM to generate similarity scores for reward calculation is a common practice in recent work [77, 20, 59] in the language domain.

The evidence-augmented reward is computed as $r_i^{\text{evid}} = r_i + r_e$ if both o_i is correct and $e_i = 1$; otherwise, $r_i^{\text{evid}} = r_i$. For each question, we compute the group reward $\mathbf{r} = \{r_i^{\text{evid}}\}_i^G$, and use it to normalize the rewards: $A_i = \frac{r_i^{\text{evid}} - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$.

With this, the policy is optimized via the clipped GRPO objective:

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

where ϵ and β are hyperparameters for clipping and KL regularization, respectively. π_θ denotes the current policy, $\pi_{\theta_{\text{old}}}$ is the prior policy used for importance sampling, and $\pi_{\theta_{\text{ref}}}$ is the reference model set to the initial checkpoint for regularization.

By incorporating an evidence-based reward signal, Visual Evidence Reward explicitly encourages models to ground their reasoning in visual context, leading to more contextually relevant responses.

Visual Evidence Generation What exactly qualifies as visual evidence? A straightforward approach might be to use general video captions [1]—but this quickly runs into the issue of granularity. Captions often miss the specific visual cues needed to answer particular questions. To resolve this, we define visual evidence in a question-dependent manner: it consists of the visual details necessary to answer a given question, which can vary significantly across tasks.

To generate such evidence, we introduce *inverted prompting*. Unlike conventional CoT prompting, where the model jointly explores reasoning paths and predicts the final answer, in inverted prompting we instead provide an external MLLM (we use Qwen-2.5-VL-72B [5]) a *(question, ground-truth answer)* pair and prompt it to produce visual evidence that *supports* this known answer. This inversion yields two key advantages. First, by fixing the answer beforehand, the model's search space is substantially reduced: it no longer needs to reason through multiple possibilities, but rather focuses solely on retrieving the *minimal, verifiable visual cues* that justify the given answer (e.g., "a red ball enters the basket"). Second, whereas standard CoT prompting can lead models to wander through generic or irrelevant reasoning before converging on an answer, our inverted prompting explicitly constrains each evidence snippet to directly explain the known answer in relation to the question. This induces an intrinsic alignment effect—any statement that is irrelevant to the question or unsupported by the video is inherently unproductive and thus discouraged under our binary evidence reward. Full prompt details are provided in the supplementary materials. Qualitative examples of the generated visual evidence used for GRPO training are shown in Figure 5.

The external MLLM is only used to generate training data in the form of question-specific visual evidence. Our policy model (Video-VER), trained on this evidence, performs inference independently and does not rely on the external VLM at test time. We recognize that MLLM-generated outputs may include speculative or hallucinated content. To mitigate this, we filter and structure the visual evidence via carefully designed prompts (see Appendix), ensuring that the extracted visual details are question-relevant and textually explicit. Empirically, this approach results in higher alignment between reasoning chains and observable video content, as demonstrated in Figure 6.

5 Experiments

Training Strategy Our model is a post-trained Qwen2.5-VL-7B [4], employing a two-stage pipeline that combines supervised fine-tuning (SFT) and reinforcement learning (RL). The process begins with SFT, where we train the model on Video-R1-COT-165k dataset [17], a dataset providing chain-of-thought (CoT) annotations, helping bootstrap the model's reasoning abilities during the cold-start phase. Following this, the model undergoes reinforcement learning using GRPO with our Visual Evidence Reward (VER). This phase uses a dataset mixture comprising Reversed-in-Time [14] and Video-R1-260k [17] samples. The RL stage is designed to move the model beyond the constraints of supervised learning, allowing it to develop more robust and adaptable reasoning patterns through exploration and self-guided optimization. We call our final post-trained model Video-VER.

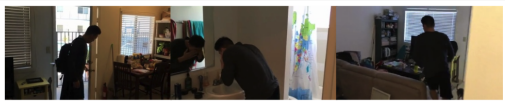
Benchmarks We extensively evaluate our model across a broad spectrum of 10 public video understanding benchmarks, covering a wide range of reasoning skills. These include comprehensive, all-around benchmarks such as MVBench [32] and Video-MME [18]; temporal reasoning benchmarks like TVBench [13], Vinoground [71], and TempCompass [38]; spatial reasoning benchmarks such as

Question: What activity does the video depict? A. Bending something so that it deforms; B. Straightening something so that it unbends.

Qwen2.5-VL-7B: The video shows a hand holding a piece of paper that appears to be bent. The hand is positioned in a way that suggests it is about to straighten the paper. The background and the context of the video support the idea that <...>. The options provided are <...>. Given the visual evidence, the action depicted in the video is the opposite of bending, which aligns with option B. ✗

Video-R1-7B: The video shows a hand holding a piece of paper that appears to be bent. The hand then seems to be manipulating the paper, possibly trying to straighten it out. Now, let's break it down: <...>. This aligns more with the action of straightening something so that it unbends. Wait, let's consider the other options. Option A suggests <...>. Oh, I see. The video clearly depicts the hand straightening the paper, <...>, aligning with option B. ✗

Video-VER-7B: The video clearly shows a hand manipulating a piece of paper. The hand initially holds the paper flat, and then it bends the paper in a specific direction. The bending action causes the paper to deform, changing its shape. This action is a clear demonstration of bending something so that it deforms. The question asks about <...>, and the options provided are <...>. However, the paper is clearly bent, and there is no evidence of it being straightened. After careful consideration, the correct answer is A. ✓



Question: What will the person do next? A. Close the door; B. Lie on the sofa; C. Open the door.

Qwen2.5-VL-7B: The person has just entered the room and is now in the bathroom. The options provided suggest that the next action could be <...> Given the context, it seems logical that the person might be preparing to leave the room, which would involve opening the door, which aligns with option C. ✗

Video-R1-7B: The person has just entered the room and has been moving around, but now they are standing near the door. What will they do next? First, let's break it down. <...> Now, they are standing near the door, <...>. Hmm, what could be their intention? They could be considering <...>, but they <...>. They could also be <...>, but they <...>. Wait, what if they <...>? It's possible that <...>. However, the door is open, <...>. Oh, I see. The most logical next step <...>, would be to close the door. <...>. Final answer is A. ✗

Video-VER-7B: The person has just entered the room and has already removed their backpack and shoes. They have walked towards the bathroom <...>. After this, they have exited the bathroom and walked back into the room. The person is now standing near the sofa, and it seems like they are preparing to sit or lie down. The question asks <...>, and the options provided are <...>. Given the context, <...>. It is most likely that they will continue their activity of sitting or lying down on the sofa. The correct answer is B. ✓

Figure 6: Comparison of reasoning traces from baselines and our Video-VER model. Notice how baseline models often include speculative or hallucinated details not grounded in the video, whereas Video-VER maintains alignment between intermediate reasoning steps and observable evidence.

VSI-Bench [63]; and knowledge-intensive datasets including Video-MMMU [24] and MMVU [76]. We also assess robustness to hallucination using dedicated benchmarks such as EventHallusion [70] and VideoHallucer [57]. Most benchmarks consist of multiple-choice questions (MCQs), with the exception of VideoMMMU and VSI-Bench, which include questions requiring numerical answers. For MMVU, we follow the protocol from [17] and use its MCQ subset to ensure robust and consistent evaluation. Subtitles are excluded in the Video-MME evaluation setting.

Implementation Details During training, the maximum number of video frames is set as 16, and increased to 32 at inference time for both our model and all baselines, unless otherwise specified. For GRPO training, we incorporate four reward components: an accuracy reward, our visual evidence reward (with weight $\alpha = 0.3$), a format reward to encourage consistent answer structure, and a length reward to promote moderately long, informative responses. We train our model with 8 NVIDIA H200 GPUs. GRPO group size G is set as 8. The number of RL iterations is set to 2,000. Further implementation details can be found in the supplementary material.

Baselines We compare our model against both proprietary and open-source video LLMs. For the proprietary model, we evaluate GPT-4o [26] and OpenAI-o3[45]. For open-source baselines, we evaluate against LongVA [72], Video-UTR [67], LLaVA-OneVision [29], Kangaroo [35], VITED [39], AoTD [48], and Qwen2.5-VL [5]. Additionally, to assess reasoning capabilities, we include two recently released video reasoning models: TinyLLaVA-Video-R1 [74] and Video-R1 [17], both of which are explicitly designed for text-based multi-step video reasoning tasks. Together, these baselines span the full spectrum of contemporary video-language systems—from large, commercially deployed models to lean, community-driven releases—ensuring that our evaluation is both representative of the strongest available competitors and informative for researchers and practitioners who rely on open tools.

5.1 Evidence-Grounded Chains Lead to Better Video Understanding

Table 1 demonstrates that Video-VER consistently surpasses existing open-source video MLLMs across a comprehensive suite of benchmarks, ranking first in 9 out of 10 evaluations, except for the VSI-Bench where Video-R1 [17] achieves the strongest performance. Video-VER shows superior results in temporally nuanced data like TempCompass (74.0%) and TVBench (52.8%), underscoring

Table 1: Comparison of Video-VER with baselines across 10 video benchmarks. Our model consistently ranks **first** or **second** overall, demonstrating the effectiveness of evidence-grounded chain-of-thought (CoT) reasoning. Notably, across most base models (e.g., Qwen2.5-VL), CoT prompting often leads to lower accuracy compared to direct answering (DA), highlighting the risk of ungrounded reasoning. In contrast, Video-VER maintains or improves performance with CoT by explicitly grounding reasoning in question-relevant visual evidence. To emphasize this, we annotate the accuracy margins (↑) between Video-VER and its base model Qwen2.5-VL-7B—both with CoT—in small offset font, drawing attention to the consistent gains enabled by our visual grounding reward.

Model	Size	Prompt	MVBench	Video-MME	VideoMMMU	MMVU	VideoHallucer	EventHallusion
GPT-4o [3]	-	DA	62.9	68.7	56.7	75.5	61.8	83.9
GPT-4o [3]	-	COT	65.1	67.0	65.0	78.9	63.4	83.6
OpenAI-o3 [45]	-	-	69.7	74.1	77.2	80.5	66.2	82.9
TinyLLaVA-Video-R1 [74]	3B	COT	49.5	46.6	-	46.9	-	-
LongVA [72]	7B	DA	-	52.6	23.9	-	-	-
Video-UTR [67]	7B	DA	58.8	52.6	-	-	-	-
LLaVA-OneVision [29]	7B	DA	57.1	57.7	33.8	49.2	34.7	61.1
VITED [39]	7B	COT	61.0	-	-	-	-	-
AoTD [48]	7B	COT	55.6	45.0	-	-	-	-
Video-R1 [17]	7B	COT	**63.9**	**59.3**	**52.3**	63.8	44.2	51.4
Kangeroo [35]	8B	DA	61.1	56.0	-	-	-	-
Qwen2.5-VL [5]	3B	DA	62.8	55.9	41.7	59.5	44.8	54.9
Qwen2.5-VL [5]	3B	COT	57.4	53.2	33.8	53.6	40.4	49.9
Qwen2.5-VL [5]	7B	DA	63.6	**59.2**	47.3	**64.2**	**51.8**	64.5
Qwen2.5-VL [5]	7B	COT	59.8	54.7	47.8	60.5	44.1	**67.3**
Video-VER (Ours)	7B	COT	**64.1** (+4.3)	**59.3** (+4.6)	**52.7** (+4.9)	**65.1** (+4.6)	**53.1** (+9.0)	**70.0** (+2.7)

Model	Size	Prompt	VSI-Bench	TempCompass	TVBench	Vinoground		
						Text	Video	Group
GPT-4o [3]	-	DA	27.8	77.8	55.1	57.6	34.4	23.8
GPT-4o [3]	-	COT	45.3	79.3	60.9	62.2	38.0	23.6
OpenAI-o3 [45]	-	-	49.3	83.4	65.9	74.4	63.0	52.8
LongVA [72]	7B	DA	-	56.9	-	-	-	-
Video-UTR [67]	7B	DA	-	59.7	-	-	-	-
LLaVA-OneVision [29]	7B	DA	32.9	67.8	47.2	42.0	28.4	12.8
AoTD [48]	7B	COT	28.8	-	-	-	-	-
Video-R1 [17]	7B	COT	**35.8**	73.2	**52.4**	34.6	24.8	9.4
Kangeroo [35]	8B	DA	-	62.5	-	-	-	-
Qwen2.5-VL [5]	3B	DA	30.1	67.3	45.2	30.2	21.2	6.0
Qwen2.5-VL [5]	3B	COT	25.4	59.8	47.2	30.8	22.6	5.8
Qwen2.5-VL [5]	7B	DA	32.3	**73.7**	52.2	**42.2**	**29.2**	**14.8**
Qwen2.5-VL [5]	7B	COT	31.4	71.3	49.9	40.6	28.0	12.8
Video-VER (Ours)	7B	COT	**34.6** (+3.2)	**74.0** (+2.7)	**52.8** (+2.9)	**42.8** (+2.2)	**28.2** (+0.2)	**14.4** (+1.6)

Table 2: Ablation study on the scalability of Video-VER across varying temporal context lengths.

Model	#Frames	MVBench	Video-MME	VideoMMMU	MMVU	VideoHal.	EventHal.	VSI-Bench	TempC.	TVBench	Vinog.
Video-VER	8	60.5	53.3	45.4	63.2	50.4	**70.5**	32.6	69.6	48.3	11.0
Video-VER	16	63.2	56.0	50.0	64.8	51.4	69.8	**35.2**	72.8	51.0	10.6
Video-VER	32	**64.1**	**59.3**	**52.7**	**65.1**	**53.1**	70.0	34.6	**74.0**	**52.8**	**14.4**

Table 3: Ablation study on prompting strategies for visual evidence generation.

Model	#Type	MVBench	Video-MME	VideoMMMU	MMVU	VideoHal.	EventHal.	VSI-Bench	TempC.	TVBench	Vinog.
Video-VER	VC	63.9	58.7	52.2	64.8	52.5	**70.3**	34.4	73.6	52.4	13.6
Video-VER	IP	**64.1**	**59.3**	**52.7**	**65.1**	**53.1**	70.0	**34.6**	**74.0**	**52.8**	**14.4**

its ability to interpret sequential information and dynamic visual content with precision. These results validate the strength of our proposed method and its generalization across diverse video-language tasks, all while utilizing a 7B parameter model with an innovative RL training strategy.

Figure 6 presents qualitative examples of the thinking chains generated by our model and baseline methods. Some raw text has been omitted for brevity. The results illustrate that the reasoning of baseline models is often distracted by speculative or hallucinated details, which are not grounded in the actual actions or the full temporal context of the video. In contrast, Video-VER better maintains alignment between intermediate reasoning steps and observable evidence, leading to the correct answer. We discuss common **failure cases** and **limitations** of our Video-VER in the appendix.

5.2 Ablation Study

Scalability with Frames Table 2 presents an ablation study assessing the scalability of Video-VER with varying numbers of input frames. The Group score is reported for the Vinoground benchmark. The results reveal a consistent trend: performance improves as more frames are provided, with the 32-frame setting achieving the best results on all benchmarks. This demonstrates that our

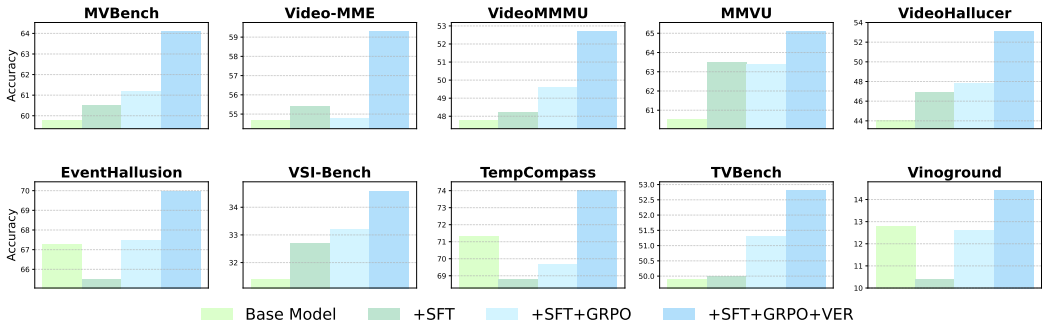


Figure 7: Ablation of post-training strategy contributions — SFT, GRPO, and VER.

method effectively leverages more frames with longer temporal context, confirming its ability to scale and generalize well across different video lengths.

Types of Visual Evidence As shown in Table 3, we evaluate two prompting strategies for generating visual evidence: our proposed Inverted Prompting (IP) and Video Captioning (VC), where the MLLM is prompted to produce a detailed caption of the video alone. The Group score is reported on the Vinoground benchmark. Our ablation reveals that IP surpasses the other approach on 9 out of 10 benchmarks, trailing only on EventHallusion. These results underscore the effectiveness of aligning visual evidence with the specific question, confirming that question-conditioned context provides greater benefit than generic video descriptions.

Ablation Study on Post-Training Strategies To investigate whether the performance gains stem from SFT, GRPO, or VER, we present detailed ablation results comparing the following configurations: the base model post-trained with SFT, the base model post-trained with SFT and GRPO, and the base model post-trained with SFT and GRPO with our VER (full model). The base model used in this study is Qwen2.5-VL-7B. In Figure 7, we show that fine-tuning the base model with SFT alone does not consistently improve performance. Adding GRPO yields moderate gains across most benchmarks. Notably, incorporating our Visual Evidence Reward (VER) on top of GRPO leads to substantial additional improvements. These results help isolate the contribution of each component and demonstrate that the gains observed in our full model are primarily driven by VER, validating its effectiveness in promoting grounded and accurate reasoning.

6 Conclusions and Future Work

Bridging the gap between the linear, symbolic structure of chains-of-thought and the inherently fuzzy, non-linear, and temporally distributed nature of video remains a central challenge. This paper has highlighted a critical failure mode in this pursuit: "*Visual Thinking Drift*", where reasoning processes, despite appearing coherent, become unmoored from actual video content. We then introduced Visual Evidence Reward (VER), a novel reinforcement learning reward mechanism specifically designed to counteract this drift. Our VER excels in closed-ended tasks like multiple-choice question answering where rule-based reward computation is effective. Extending this framework to open-ended tasks, such as free-form QA, and thereby ensuring that verbosity is replaced by genuinely grounded video intelligence, presents an exciting and crucial avenue for future work.

Societal Impact The advancements in video reasoning presented in this paper, particularly the Visual Evidence Reward framework designed to mitigate *Visual Thinking Drift* and enhance the grounding capabilities of models like Video-VER, offer significant societal potential, encompassing both promising benefits and important ethical considerations. More accurate, reliable, and interpretable video understanding systems can yield substantial advantages across diverse domains, including improved accessibility for individuals with visual impairments through richer descriptions of visual media, enhanced educational tools that can better analyze instructional content, and more dependable automated content analysis due to a reduction in model-generated hallucinations. On the other hand, such powerful video reasoning capabilities may also introduce risks related to privacy infringement or the amplification of bias in data-driven systems, underscoring the importance of responsible development and deployment practices.

Acknowledgements This research has been supported by computing support on the Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin. UT Austin is supported by NSF Grants 2019844, 2112471, AF 1901292, CNS 2148141, Tripods CCF 1934932, Tripods 2217069, NSF AI Institute for Foundations of Machine Learning (IFML) 2019844, the Texas Advanced Computing Center (TACC) and research gifts by Western Digital, Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program, UT Austin Machine Learning Lab (MLL), Cisco and the Stanly P. Finch Centennial Professorship in Engineering.

We thank the anonymous reviewers for their constructive feedback. We are also grateful to Yue Zhao for valuable discussions on MLLM-based video reasoning, and to Kumar Ashutosh and Sagnik Majumder for their insightful feedback on the initial model design.

References

[1] Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi, Farhad Pourpanah, Daniel McDuff, Mohammad Ghavamzadeh, Shuicheng Yan, Abduallah Mohamed, Abbas Khosravi, Erik Cambria, et al. A review of deep learning for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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. *arXiv preprint arXiv:2502.13923*, 2025.

[6] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

[9] Zhe Chen, Jiannan Wu, Wenhai 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.

[10] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Hanoona Rasheed, Peize Sun, Po-Yao Huang, Daniel Bolya, Suyog Jain, Miguel Martin, Huiyu Wang, Nikhila Ravi, Shashank Jain, Temmy Stark, Shane Moon, Babak Damavandi, Vivian Lee, Andrew Westbury, Salman Khan, Philipp Krähenbühl, Piotr Dollár, Lorenzo Torresani, Kristen Grauman, and Christoph Feichtenhofer. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv:2504.13180*, 2025.

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[13] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Tvbench: Redesigning video-language evaluation, 2024.

[14] Yang Du, Yuqi Liu, and Qin Jin. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *Proceedings of the 32th ACM International Conference on Multimedia*, page 5260–5269.

[15] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024.

[16] 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. *arXiv preprint arXiv:2501.03230*, 2024.

[17] 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 preprint arXiv:2503.21776*, 2025.

[18] Chaoyou Fu, Yuhan Dai, Yondong 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. *arXiv preprint arXiv:2405.21075*, 2024.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[20] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[22] 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. *arXiv preprint arXiv:2411.14794*, 2024.

[23] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

[24] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. 2025.

[25] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024.

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[27] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[28] Hector J Levesque. Knowledge representation and reasoning. In *Readings in Artificial Intelligence and Databases*, pages 35–51. Elsevier, 1989.

[29] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[30] Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. *arXiv preprint arXiv:2501.01926*, 2025.

[31] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[32] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[33] 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.

[34] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024.

[35] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024.

[36] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*, 2024.

[37] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025.

[38] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

[39] Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, and Tushar Nagarajan. Vited: Video temporal evidence distillation. *arXiv preprint arXiv:2503.12855*, 2025.

[40] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 3, 2024.

[41] Jianguo Mao, Wenbin Jiang, Xiangdong Wang, Zhifan Feng, Yajuan Lyu, Hong Liu, and Yong Zhu. Dynamic multistep reasoning based on video scene graph for video question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3894–3904, 2022.

[42] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[43] Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, et al. Minerva: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*, 2025.

[44] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. 2021.

[45] OpenAI. Openai o3 and o4-mini system card. https://openai.com/index/o3-o4-mini-system-card/, 2025. Accessed: 2025-10-21.

[46] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2402.13950*, 2024.

[47] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[48] Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. *arXiv preprint arXiv:2412.01694*, 2024.

[49] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134, 1999.

[50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[51] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[52] Open Thoughts Team. Open Thoughts, January 2025.

[53] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

[54] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*, 2024.

[55] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024.

[56] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[57] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arxiv*, 2024.

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[59] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. *arXiv preprint arXiv:2309.11489*, 2023.

[60] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*, 2024.

[61] Zihui Xue, Joungbin An, Xitong Yang, and Kristen Grauman. Progress-aware video frame captioning. In *CVPR*, 2025.

[62] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. In *NeurIPS*, 2025.

[63] Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024.

[64] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[65] Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*, 2023.

[66] Chris Yi Peng et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought, 2025.

[67] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025.

[68] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023.

[69] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[70] Jiacheng Zhang, Jiao Yang, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv: 2409.16597*, 2024.

[71] Jianrui Zhang, Cai Mu, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv*, 2024.

[72] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[73] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.

[74] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025.

[75] 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, April 2024.

[76] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025.

[77] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[78] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline the key objectives and outcomes of the research, including the problem being addressed, the proposed approach or methodology, and the significance of the findings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discussed the limitations of our work in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Not applicable, as this paper does not include any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the necessary information to reproduce the main experimental results in the Experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data will be released.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper includes comprehensive information on key training configurations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We conducted multi-run evaluations for a subset of the experiments and observed no significant variance in the results. Due to the high computational cost—approximately 384 H200 GPU hours per run—it was not feasible to perform multiple runs for all experiments. Nonetheless, the stability observed in the sampled multi-run tests supports the reliability of the reported results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to the Experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that the research conducted in this paper adheres to its guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the discussion provided in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We take responsible release seriously and ensure that any code or evaluation data we release avoids including sensitive content or misuse-enabling functionality. Our work is intended for academic benchmarking and insight into model behavior.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Any pretrained models referenced are publicly available and subject to their original licenses and usage policies.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will provide guideline documentation as part of the release.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[NA]

    Justification: This study does not involve any human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This study does not involve any human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for writing, editing, or formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Contents

A Supplementary Video

For a dynamic and visual exploration of our method, we encourage readers to watch the supplementary video available at https://vision.cs.utexas.edu/projects/video-ver/. It provides an intuitive walkthrough of how our proposed Visual Evidence Reward (VER) operates within the GRPO framework and presents further qualitative examples that highlight its effectiveness in grounding video reasoning.

B Bayesian Perspective on Visual Thinking Drift

We provide a more nuanced explanation of the root cause of the performance degradation caused by Visual Thinking Drift, particularly as it pertains to the unique challenges in video-based tasks.

Consider a question q, condition $\mathbf{v}$ (which represents either the video features in video LLMs or the "prefilling text" in text-only LLMs), reasoning tokens $c_{1:T}$ and final answer a, the joint probability under an autoregressive video-LLM is

$$p(c_{1:T}, a \mid q, \mathbf{v}) = p\big(a \mid c_{1:T}, q, \mathbf{v}\big) \prod_{t=1}^{T} p\big(c_t \mid c_{<t}, q, \mathbf{v}\big). \tag{1}$$

We now derive an explicit form for each conditional $p\big(c_t \mid c_{<t}, q, \mathbf{v}\big)$. In video LLMs, at step t the decoder's last hidden vector can be written as

$$\mathbf{h}_t = \underbrace{\mathbf{h}_t^{(\text{lang})}}_{\text{from previous text tokens}} + \underbrace{A_{\text{vis}}\big(\mathbf{r}_t^{(\text{vis})}\big)}_{\text{from visual adapter}}, \tag{2}$$

where $A_{\text{vis}} \in \mathbb{R}^{d \times d_{\text{vis}}}$ is the vision adapter (either a linear layer or MLP), and $\mathbf{r}_t^{(\text{vis})}$ is the visual embedding.

During decoding, every token passes through the same vocabulary projection $W_{\text{out}} \in \mathbb{R}^{d \times |\mathcal{V}|}$. For a candidate token w with one-hot encoding $\mathbf{e}_w$, the logit is computed as $z_w = \mathbf{h}_t^\top W_{\text{out}} \mathbf{e}_w$.

Substituting into Equation (2),

$$\begin{aligned} z_w &= (\mathbf{h}_t^{(\text{lang})})^\top W_{\text{out}} \mathbf{e}_w + \big(A_{\text{vis}} \mathbf{r}_t^{(\text{vis})}\big)^\top W_{\text{out}} \mathbf{e}_w \\ &= \mathbf{h}_{c_{<t}}^\top \underbrace{W_{\text{out}}}_{W_{\text{lang}}} \mathbf{e}_w + \mathbf{h}_{\mathbf{v}}^\top \underbrace{\big(A_{\text{vis}}^\top W_{\text{out}}\big)}_{W_{\text{vis}}} \mathbf{e}_w, \end{aligned} \tag{3}$$

where we define

$$\mathbf{h}_{c_{<t}} \equiv \mathbf{h}_t^{(\text{lang})}, \qquad \mathbf{h}_{\mathbf{v}} \equiv \mathbf{r}_t^{(\text{vis})}, \qquad W_{\text{vis}} = W_{\text{out}} A_{\text{vis}}.$$

The conditional distribution is given by a softmax over all tokens w:

$$p(c_t = w \mid c_{<t}, q, \mathbf{v}) = \frac{\exp(z_w)}{\sum_{w' \in \mathcal{V}} \exp(z_{w'})}. \tag{4}$$

Removing the common denominator yields the unnormalised expression:

$$p\big(c_t \mid c_{<t}, q, \mathbf{v}\big) \propto \exp\Big(\underbrace{\mathbf{h}_{c_{<t}}^\top W_{\text{lang}}}_{\text{language prior}} + \underbrace{\mathbf{h}_{\mathbf{v}}^\top W_{\text{vis}}}_{\text{visual likelihood}}\Big). \tag{5}$$

Here $W_{\text{lang}} \equiv W_{\text{out}}$ and $W_{\text{vis}} \equiv W_{\text{out}} A_{\text{vis}}$.

In Section 3.2, we noted that the dominance of language priors in video-LLMs stems from asymmetries in model initialization and data scale. These asymmetries lead to a substantially smaller visual projection norm, causing language signals to overpower visual inputs during decoding—formally,

$$\|W_{\text{vis}}\|_F \ll \|W_{\text{lang}}\|_F.$$

In contrast, for text-only LLMs, there is no visual term $A_{\text{vis}}(\mathbf{r}_t^{(\text{vis})})$ due to the absence of a vision encoder. As a result, the same "thinking drift" mechanism does not apply in purely textual settings.

To further substantiate this point, we now provide additional theoretical evidence supporting the presence of this asymmetry—specifically, how the dominance of the language component underlies the phenomenon we refer to as *visual thinking drift*.

A Simple Closed-form Derivation We now explain why fewer gradient updates to the vision encoder naturally lead to a smaller Frobenius norm for the visual projection matrix, compared to that of the language model.

Assume a trainable matrix evolving under stochastic gradient descent (SGD) as

$$W \leftarrow W^{(0)} + \sum_{t=1}^{N} \Delta_t,$$

where the updates Δ_t have zero mean and second moment $\mathbb{E}\big[\|\Delta_t\|_F^2\big] = \sigma^2$. Then the expected squared Frobenius norm becomes

$$\mathbb{E}\big[\|W\|_F^2\big] = \|W^{(0)}\|_F^2 + N\sigma^2, \quad \Longrightarrow \quad \mathbb{E}\big[\|W\|_F\big] \propto \sqrt{N}.$$

This shows that the expected Frobenius norm grows with the square root of the number of updates N.

In modern video-LLMs, the language pathway typically consists of a massive pre-trained backbone trained on trillions of text tokens (e.g., $\sim$18T for Qwen 2.5). In contrast, the visual pathway often involves a newly introduced adapter module trained on significantly fewer multimodal examples—on the order of tens of billions of paired tokens (e.g., $\sim$40B for Flamingo). As a result, the number of gradient updates N for the visual projection is much smaller, leading naturally to the observed asymmetry:

$$\|W_{\text{vis}}\|_F \ll \|W_{\text{lang}}\|_F.$$

C Self-consistency Decoding for Video Reasoning

Motivated by the decoding strategy self-consistency proposed in [56], which samples a diverse set of reasoning paths instead of relying solely on greedy decoding, and then selects the most consistent answer by marginalizing over the sampled paths, we explore its implications for video reasoning tasks. The intuition behind self-consistency is that complex reasoning problems often allow for multiple valid reasoning trajectories that converge on a unique correct answer.

However, our focus is on video reasoning tasks, which generally exhibit lower reasoning complexity than language-only tasks such as mathematical problem solving or code generation. Moreover, the diversity of reasoning paths in video-based tasks tends to be more constrained due to the fixed visual context and limited temporal narrative. For example, while a math problem might allow several logical formulations or decompositions, a video clip typically presents a specific sequence of events that restricts interpretive variation.

Despite this, we observe that simple majority voting over 20 responses generated using Chain-of-Thought (CoT) prompting (with the same model) significantly boosts accuracy across all models in most scenarios. See Figure 8. In particular, this indicates that the reasoning traces for video tasks are often unstable, and greedy decoding is more prone to getting trapped by the *visual thinking drift*—the phenomenon discussed in Section 3.2, where subtle ambiguities or misinterpretations in the visual context derail the reasoning process. By aggregating multiple responses, self-consistency helps to smooth out these drifts and converge on more robust answers.

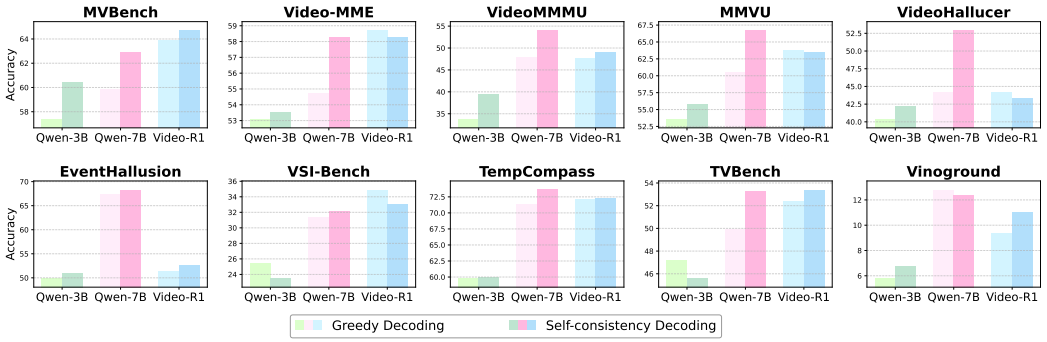


Figure 8: Sampling multiple independent CoT responses per question and aggregating them via majority voting yields a clear accuracy improvement—indicating that reasoning traces are often stochastic rather than dependable. Each chart is a different popular video benchmark.

D Implementation Details

Training and Testing Configurations For evaluation, we set the temperature to 0.01 for all baseline models as well as our model. During training, the maximum video token budget is set to $128 \times 28 \times 28$ pixels. During testing, this is increased to $256 \times 28 \times 28$ pixels. The sampling rate is set to 2.0 FPS across all benchmarks, except for Vinoground, which uses 4 FPS. To manage API costs, the maximum number of frames used by GPT-4o is limited to 16 frames per video, except for TempCompass, where only 8 frames are used.

The results for VSI-Bench reported in Table 1 represent the average accuracy across both multiple-choice and regression-based tasks. For the baseline model Video-R1, we adopt the reported results of MVBench, Video-MME, VideoMMMU, MMVU, VSI-Bench, and TempCompass under the 32-frame setting as specified in their respective papers. For the remaining four benchmarks—VideoHallucer, EventHallusion, TVBench, and Vinoground—we conduct our own evaluations due to the absence of publicly available results. In Figure 2 and Figure 8, to ensure a fair and consistent comparison across all prompting and decoding strategies, we re-evaluate Video-R1 on all benchmarks in our experimental setting, including those with reported numbers, wherever applicable.

Length Reward To encourage deeper reasoning without excessive verbosity, we apply a length reward targeting response lengths in the range of 320 to 512 tokens.

Direct Answer Prompt

{**QUESTION**}
Please provide your answer within the <answer> </answer> tags.

Chain-of-Thought Prompt

{**QUESTION**}
Please think about this question as if you were a human pondering deeply.

Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions.

It's encouraged to include self-reflection or verification in the reasoning process.

Provide your detailed reasoning between the <think> and </think> tags, and then give your final answer between the <answer> and </answer> tags.

Figure 9: Prompts used for direct response generation and reasoning-driven generation (chain-of-thought). The CoT prompt is borrowed from Video-R1 [17].

Visual Evidence Generation Prompt

{**QUESTION**}
The answer is {**ANSWER**}.
Please briefly cite the visual details—objects, actions, relations, and temporal order—that support this answer. Enclose the evidence within the <evidence> and </evidence> tags.

Figure 10: Inverted Prompting used for generating visual evidence annotations.

LLM Judge We utilize Llama-3.1-70B-Instruct [19] as our LLM-based judge. It is prompted to produce a binary label—1 for successfully referencing the visual evidence, and 0 for failing to do so.

Prompt Details All prompts used in our training and evaluation are illustrated in Figures 9 and 10.

E LLM-Based Judge for Visual Evidence Grounding

Evaluating whether a model's reasoning references the correct visual evidence is inherently a semantic task that goes beyond exact string or token overlap. It requires assessing whether the generated rationale mentions visual facts that are relevant, specific, and consistent with what is shown in the video. To enable this, we adopt an auxiliary LLM-based judge to compute a binary reward—assigning 1 if the rationale includes grounded, question-relevant visual details, and 0 otherwise.

This design aligns with common practice in recent reinforcement learning and reward modeling studies, where LLMs are used to produce reward signals for complex, fuzzy objectives such as factuality, helpfulness, or alignment [77, 20, 59]. Unlike token-level metrics, the LLM judge offers flexible, high-level reasoning about textual similarity and visual grounding. While this introduces some approximation and risk of bias, it is essential for enabling scalable supervision at training time where human annotation is impractical. To reduce their influence we use temperature-0 decoding for more deterministic decoding and use the carefully designed prompt shown in Figure 11. In Figure 13, we present qualitative examples where the LLM-assigned rewards reflect accurate judgments of visual grounding, demonstrating the judge's effectiveness in identifying whether reasoning is properly anchored in the video content.

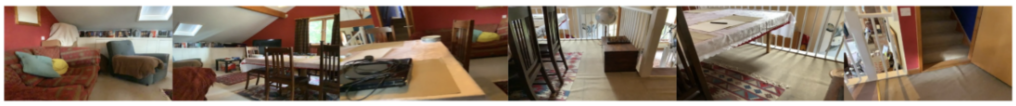
Prompt for LLM-Judge

You are an expert in evaluating whether the reasoning trace of a video reasoning model is grounded in the visual evidence. You will be provided with a reasoning trace and a description of the visual evidence. Determine whether the reasoning process uses facts from the visual evidence to arrive at the final answer.

Visual Evidence: {**VISUAL EVIDENCE**}
Thinking Trace: {**THINKING TRACE**}

Output only a single numeric value: **1** if the reasoning uses the visual evidence, **0** if it does not.

Figure 11: The prompt used for LLM Judge, producing reward value 0 or 1.



Question: How many chair(s) are in this room? Solution: 9

Qwen2.5-VL-7B: Let me think through this carefully. The room has a dining table with chairs around it. There are two chairs visible in the dining area. Additionally, there is a single chair in the living area, which is not part of the dining table. To conclude, the dining area has two chairs, and the living area has one chair. Adding these together gives us a total of **3** chairs in the room. ✗

Video-R1-7B: Let me think about this carefully. The question is asking about the number of chairs in the room. Now, I need to recall the details of the room from the video. First, I'll break down the room's layout. There's a dining table with chairs around it, and I can see a couple of chairs near the table. Then, there's a couch with a couple of cushions, and I don't see any chairs there. Wait, let's focus on the dining area. There's a table with a white tablecloth, and around it, there are chairs. I can clearly see four chairs around the table. Hmm, I need to double-check if there are any other chairs in the room. I don't see any other chairs in the living area or the kitchen area. The room is quite spacious, and I've gone through all the visible areas, and I'm confident there are no other chairs. Oh, I see. The room has four chairs around the dining table. There are no other chairs in the room. Let's break it down one more time to be sure. The dining area has four chairs, and the rest of the room doesn't have any additional chairs. After this thorough analysis, I'm confident that the room has exactly **4** chairs. Let me verify one last time. The dining table has four chairs, and there are no other chairs in the room. Yes, that's correct. So, the number of chairs in the room is four. ✗

Video-VER-7B: Let me think. The video showcases a cozy, multi-purpose room with a warm and inviting atmosphere. The room features a dining area with a wooden table and chairs, a comfortable sofa adorned with cushions, and a small table with books and decorative items. The walls are painted in a soft beige color, complemented by wooden beams and a skylight that allows natural light to flood the space. The room also includes a staircase leading to an upper level, adding to its charm. The overall ambiance is warm and welcoming, making it an ideal space for relaxation and socializing. Let's break down the video content to identify the number of chairs present in the room. The video starts with a view of the dining area, where a wooden table is set with plates and utensils. There are chairs around the table, but the exact number isn't immediately clear. As the camera pans around the room, it reveals more details. The dining area is adjacent to a living area with a sofa and a small table. The camera then moves to the staircase, which leads to an upper level. The staircase area has a few chairs placed near it, possibly for resting or reading. The camera continues to pan, showing the entire room, and it becomes evident that there are a total of 4 chairs in the room. The chairs are distributed around the dining table and near the staircase. Now, let's verify the count. The dining table has 4 chairs, and there are 2 additional chairs near the staircase. This brings the total to **6** chairs. ✗

Figure 12: A common failure case in which both the baseline models and our method produce incorrect reasoning traces due to missing visual context, caused by incomplete frame sampling.

F Visual Evidence Generation: Bootstrapping Grounded Reasoning

Generating high-quality, question-specific visual evidence is crucial for anchoring Chain-of-Thought (CoT) reasoning in video understanding. This task demands identifying precise visual cues—objects, actions, and temporal events—that directly justify a given answer. Given the current limitations of most video-language models (MLLMs) in reliably producing such granular evidence without explicit supervision, we employ a strong external MLLM, Qwen2.5-VL-72B, as an **offline generator** during training. Our goal is to bootstrap a lightweight, yet effective, reward signal that explicitly encourages reasoning traces to be grounded in observable visual content, thereby circumventing the need for expensive human annotation.

Potential Concerns Introducing an external "teacher" MLLM for evidence generation naturally raises valid concerns regarding noise, potential circularity in the training process, and the risk of hallucinated details. We've implemented several strategies to rigorously mitigate these issues.

First, and crucially, the external MLLM is utilized *solely offline* to create question-specific visual evidence for training. Our policy model, Video-VER, learns from these generated outputs but operates

Table 4: Comparison of models on NExT-QA and EgoSchema benchmarks.

Category	Model	NExT-QA	EgoSchema
Agentic Workflow	SeViLA [68]	73.8	-
Agentic Workflow	VideoAgent [55]	71.3	54.1
Agentic Workflow	MoReVQA [42]	69.2	51.7
MLLM End-to-End Training	Video-VER (Ours)	81.0	66.2

entirely independently at inference time. This ensures there's no direct feedback loop or reliance on the teacher's rationale during deployment, preserving the integrity of our trained model.

Second, the reward signal derived from the generated visual evidence is intentionally *binary*. We don't demand exhaustive or perfectly nuanced evidence; instead, the judge model merely verifies if the generated reasoning trace cites *any* verifiable visual fact relevant to the answer. This design choice significantly attenuates the impact of occasional hallucinations or minor inaccuracies from the teacher model: incorrect or missing evidence simply results in a zero reward, rather than actively pushing the policy towards an erroneous trace. This robustness ensures that the reward signal guides the model toward *grounded reasoning*, rather than penalizing minor deviations.

Third, a core innovation in our visual evidence generation strategy lies in our *inverted prompting approach*. Unlike conventional CoT, where the model simultaneously explores reasoning steps and the final answer, we feed the external MLLM the *(question, ground-truth answer)* pair and instruct it to generate visual evidence that *supports* this already-known answer. This inversion offers two benefits. By fixing the answer upfront, the model's task is narrowed considerably. It no longer has to navigate the full reasoning space; its sole objective is to retrieve the *minimal, verifiable visual facts* that logically lead to the given answer (e.g., "a red ball enters the basket"). Furthermore, in typical CoT, models can sometimes drift into generic narratives before settling on an answer. Our inverted prompt structure forces each generated evidence snippet to directly explain the already-known answer to the specific question asked. This creates an intrinsic alignment pressure: any statement irrelevant to the question or unsupported by the video becomes effectively useless, and thus discouraged by our binary evidence reward. The model is incentivized to produce *highly relevant and visually verifiable facts*, directly addressing the "Visual Thinking Drift" problem by ensuring reasoning is tethered to pertinent visual cues. A potential concern is that providing the question text might allow the MLLM to generate evidence based on linguistic cues, bypassing direct pixel analysis. However, our empirical results in Table 3 address this: generating visual evidence conditioned on the question text significantly outperforms using generic video captions. This suggests that the question text's primary contribution is not an increase in hallucination; rather, it provides crucial, task-specific context that guides the MLLM towards more relevant visual features, thereby enhancing overall performance.

Put differently, a standard CoT prompt induces the distribution $p(c_{1:T}, a \mid q, \mathbf{v})$, where both the reasoning chain and the answer are uncertain. Our evidence prompt, by contrast, conditions on the correct answer and samples from $p(e_{1:K} \mid q, a, \mathbf{v})$. This constitutes a far lower-entropy target that inherently prioritizes *visual grounding over linguistic priors*. This fundamental structural difference in how the evidence is generated explains why our teacher model can reliably produce visual evidence *without* itself needing the very reinforcement signal we are about to learn.

Figure 13 provides further samples of our generated visual evidence, demonstrating how our *inverted prompting* (IP) approach yields highly relevant and specific visual cues that directly support the answers, effectively anchoring the reasoning process.

G Agentic Workflows vs. MLLM End-to-End Training

The performance of agentic workflow approaches [68, 55, 42] is often constrained by the capabilities of their component visual secialist models (e.g., key-frame selectors or object detectors). In contrast, end-to-end training leverages a single multimodal LLM trained on large-scale vision-language data, with the vision encoder handling perception and the language model handling reasoning—enabling a more unified approach to video understanding. Our VER model follows this end-to-end paradigm. As a reference, we include additional evaluations of recent agentic workflow methods [68, 55, 42], which approach video understanding by decomposing the task into sub-problems—such as key-frame selection or temporal grounding—and addressing each with specialized models. Across benchmarks, our 7B Video-VER model demonstrates superior overall performance compared to these baselines.

Nonetheless, agentic workflows offer some flexibility at test time. Their ability to decompose tasks and incorporate external visual models may provide advantages in specific scenarios.

H Failure Case

We illustrate a common failure case where both the baseline models and our Video-VER model generate incorrect reasoning traces. This occurs due to incomplete frame sampling, which omits critical visual context from the video tokens needed to answer the question accurately. See Figure 12.

I More Qualitative Examples

As shown in Figure 14 and 15, we present additional qualitative examples of reasoning chains produced by our model alongside the baseline models Qwen2.5-VL-7B and Video-R1-7B. These examples further demonstrate that baseline models frequently rely on speculative or hallucinated details, often misaligned with the actual actions or broader temporal context of the video. In contrast, Video-VER consistently grounds its intermediate reasoning steps in observable evidence, resulting in more accurate answers.

J Limitations

Our Visual Evidence Reward framework and Video-VER model significantly advance grounded video reasoning. As with any research, there are aspects that provide context for our findings and suggest avenues for future exploration:

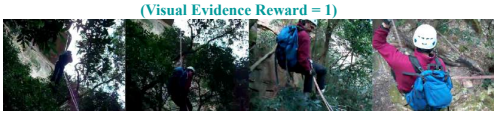
First, the performance of video reasoning systems, including Video-VER, is closely tied to the nature of the input video and how it is processed. Our current investigations primarily focus on videos of moderate length. This focus considers the current ability of many state-of-the-art vision encoders to effectively process very long video sequences, and the common MLLM design where the language model reasons using visual tokens that are expected to contain all necessary information for the task. As a result, extending robust, fine-grained reasoning to scenarios with very long videos—especially those containing sparse critical information or complex temporal patterns—remains an important direction for future work, which will benefit from advances in long-sequence video encoding. Furthermore, the quality of the visual representation itself significantly influences performance, irrespective of video length. As highlighted in our failure case analysis (Figure 12), if essential visual details are missed, for instance, due to frame sampling choices or limitations in feature extraction, the reasoning process can be compromised. Our Visual Evidence Reward method ensures that reasoning is firmly based on the *available* visual data, and its effectiveness is therefore complemented by ongoing improvements in dynamic frame selection and visual feature extraction techniques.

Second, the LLM-based reward signal generation is a key component of our approach. We utilize Llama-3.1-70B-Instruct with robust prompting strategies and temperature-0 decoding to ensure reliable visual grounding assessment. The quality of the reward signal is inherently connected to the chosen LLM's capabilities. As with similar LLM-based evaluation paradigms in the field, ongoing research into optimizing these judge models and exploring alternative evaluation mechanisms will further enhance the precision and scalability of such reward mechanisms.

Third, regarding the generation of visual evidence for training, our approach employs a powerful external MLLM (Qwen2.5-VL-72B) in an offline manner. Our inventive inverted prompting strategy and binary reward design are specifically configured to maximize the relevance of this evidence and ensure robustness against potential imperfections from the teacher model. While this bootstrapping method proves effective for our training purposes, continued advancements in the capabilities of MLLMs to generate high-fidelity, nuanced visual descriptions will naturally offer opportunities to refine the training data for such reward mechanisms even further.

We view this work as a significant step towards overcoming visual thinking drift and achieving more robust, visually grounded video intelligence. Future directions include extending our Visual Evidence Reward framework to effectively process long-form videos (potentially incorporating advanced visual information aggregation techniques) and exploring its integration with even more complex multi-step reasoning tasks that demand deeply verifiable and contextually rich thought processes.

(Visual Evidence Reward = 1)



Question: What does the person rely on for support while descending the cliffside?

Options:
A. A ladder
B. A harness
C. A rope
D. A bridge

Visual Evidence: The person is using a rope for support while descending the cliffside. This is evident from the visible rope attached to the individual, which is likely part of a climbing or rappelling setup. The person is also wearing a harness, which is a safety device used in climbing to secure the rope and prevent falls. The presence of the rope and harness, along with the individual's posture and the context of the activity, strongly supports the answer that the person relies on a rope for support.
Model Response: Let me think about this. The climber is clearly using a rope for support while descending the cliffside. The rope is visible in the video, and it is being used to control the descent, ensuring safety and stability. The climber's posture and the way they are holding onto the rope indicate that they are relying on it for support. There is no mention or indication of a ladder, a harness, or a bridge in the video. The rope is the only tool that can be seen and used for this purpose.

Question: Where is the person?

Options:
A. Indoors in the living room or bedroom
B. Outdoors
C. Indoors in the kitchen

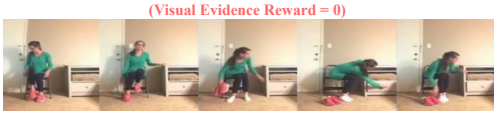
Visual Evidence: A person is preparing a mixture in a kitchen setting. They are cracking an egg into a yellow bowl, followed by adding a spoonful of sugar and a carton of lactose-free milk. The countertop is cluttered with various kitchen items, including a container of creamer and a box of chocolate. The person is wearing a sleeveless shirt with a colorful design.
Model Response: Let me think about this. The video depicts a person engaged in a cooking or baking activity, which is typically done indoors in a kitchen. The presence of kitchen items such as a bowl, sugar, milk, and creamer, along with the person's attire and the overall setting, strongly suggest that this is taking place indoors in a kitchen. The living room or bedroom would not typically have the necessary items for this activity, and the outdoor setting is unlikely given the indoor kitchen items and the person's attire.

Question: Why did the woman pull the leash when the dog is walking towards the water at the end? Options:

Options:
A. calm it down.
B. to keep the dogs running away.
C. balance dog.
D. take him away from the water.
E. bring dog to walk.

Visual Evidence: A woman is walking her dog along a lakeside path. The dog is on a leash and appears to be pulling slightly towards the water. The woman pulls the leash towards herself as the dog approaches the water. This action suggests she is trying to guide the dog away from the water, possibly to prevent it from entering or to ensure it stays on the path. The woman is walking at a steady pace, and the dog is following her lead.
Model Response: Let me think about this. The woman is likely trying to guide the dog away from the water. Dogs can sometimes be drawn to water, and the woman might be concerned about the dog getting too close to the water, possibly for safety reasons or because she wants to avoid the dog getting wet or dirty. The action of pulling the leash suggests she is actively trying to control the dog's movement.

(Visual Evidence Reward = 0)



Question: Which object would the person take next?

Options:
A. The shoe.
B. The book.
C. The towel.
D. The phone/camera.

Visual Evidence: The video shows the person removing both shoes, and the next step would be to put them away or put on the other shoe. The shoe is located on the floor near the person, making it easily accessible for the next action.
Model Response: Let me think about this. The person has already taken the shoe and placed it on the floor. They are now standing up and appear to be looking around the room. Given the options, the most logical next step would be to pick up an object that is commonly used in a room like this. The book seems like a plausible choice as it is often found on nightstands or dressers. The towel and phone/camera are less likely to be the next item of interest.

Question: The person interacts with an electric device among other objects. Is the electric device on at any point?

Options:
A. no
B. yes
C. I don't know

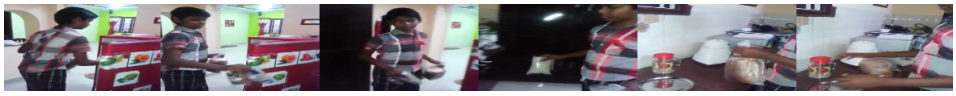
Visual Evidence: The electric device appears to be on at some point during the video. This can be inferred from the person's hand movements and the visible light on the device. The hand is shown holding the device in a way that suggests it is in operation, and there is a noticeable light on the device, which is typically an indicator that it is powered on. Additionally, the person's actions, such as adjusting the device and pointing towards it, suggest that it is functioning and being used actively.
Model Response: Let me think about this. The video shows a person holding an electric device, which appears to be a hair clipper, and a small object that could be a comb. The person is examining both items closely, and there is no indication that the electric device is turned on. The setting is a kitchen, and the person is not interacting with the device in a way that suggests it is in use. The focus is on the objects and the person's hands, not on the device's operation.

Question: Which object did the person take after they lied on the bed?

Options:
A. The phone/camera.
B. The towel.
C. The broom.
D. The paper/notebook.

Visual Evidence: After lying on the bed, the person moved to a desk and picked up a paper/notebook. This is evident from the sequence of actions: first, the person is seen lying on the bed with a laptop, then they get up and walk to the desk where they pick up a paper/notebook. The desk is cluttered with various items, including papers, a clock, and other office supplies, which suggests a workspace environment.
Model Response: Let me think about this. The question asks which object the person took after lying on the bed. The video shows the woman lying on the bed and then getting up to go to the desk. The options provided are: A. The phone/camera. B. The towel. C. The broom. D. The paper/notebook. The woman is seen sitting at the desk. The most logical object to take after lying on the bed and moving to the desk would be something they need to work on or use for their task. Therefore, the object is likely the phone or camera.

Figure 13: Visual evidence generated through our *inverted prompting* approach, as well as the evidence reward values produced by the LLM Judge. These examples illustrate the specific and relevant visual details that directly support answers and anchor the reasoning process. Key visual evidence is underlined. Green highlighting indicates successful reference to visual evidence, while red indicates a failure to do so.



Question: What did the person do first? Options: A. Closed the closet/cabinet. B. Put down the food.

Qwen2.5-VL-7B: Let me think. The person is seen standing in front of a refrigerator, which suggests they might have been retrieving something from it. However, the video then cuts to the person in a different location, which could be a kitchen, where they are handling a bottle and a jar. This implies that the initial action of retrieving something from the refrigerator might not be the first action in the sequence shown in the video. The answer is B. ✗

Video-R1-7B: Let me think about this for a moment. The question is asking what the person did first, and we have two options to consider: A. Closed the closet/cabinet, and B. Put down the food. First, let's break it down and analyze the sequence of actions in the video. The person is seen interacting with various objects in the kitchen, and there's a clear timeline of events. Wait, I need to focus on the initial action. The person was initially standing near the refrigerator, and then they moved towards the counter. They picked up a bottle and a jar, and then they started to prepare something. Hmm, it seems like the person first interacted with the refrigerator, <...>. However, the question is specifically about <...>. Oh, I see. The person first interacted with the refrigerator, and then they moved towards the counter. They didn't close the closet/cabinet immediately after entering the kitchen. They were more focused on the refrigerator and the counter. <...>. Therefore, the first action was related to the refrigerator. So, the person first interacted with the refrigerator, which implies they put down the food after taking something out or putting something in the refrigerator. Therefore, the correct answer is B. ✗

Video-VER-7B: Let me think about this for a moment. The person in the video is seen interacting with various objects in a kitchen setting. First, they appear to be reaching into a cabinet or closet, which suggests they are retrieving something. Then, they seem to be handling a bottle, possibly pouring its contents into a container. After that, they are seen holding a piece of food, which they appear to be preparing or examining. The question asks what the person did first, and the options provided are A. Closed the closet/cabinet and B. Put down the food. To answer this, I need to carefully analyze the sequence of actions in the video. First, the person reaches into a cabinet or closet, indicating they are retrieving something. Second, they then handle a bottle, which implies they are using it for some purpose. Finally, they are seen holding a piece of food, which they seem to be preparing or examining. Given this sequence, the first action the person took was reaching into the cabinet or closet to retrieve something. This aligns with option A. Hmm, let's verify this. The person did not put down any food at the beginning; they were still in the process of retrieving it from the cabinet. This confirms that the correct answer is A. ✓

Question: Does the video show a complete motorcycle? A. yes B. no

Qwen2.5-VL-7B: Let me think about this. The image shows a motorcycle, but it appears to be a close-up shot, focusing on the front part of the bike. The front wheel, headlight, and part of the body are visible, but the rear wheel and other parts of the motorcycle are not shown. This suggests that the video might not show a complete motorcycle. The answer is B. ✗

Video-R1-7B: Let me think about this for a moment. The question is asking <...>. Now, let's break it down. The video seems to focus on a motorcycle, but it's important to consider what exactly is shown. First, the video shows a motorcycle in a garage, and it's illuminated by a bright light. The motorcycle is clearly visible, and it appears to be a scooter, given its design and features. The lighting highlights the details of the motorcycle, such as the headlights, the bodywork, and the wheels. However, the question is <...>. A complete motorcycle would typically include all parts and features that make it a fully functional vehicle. In this video, the motorcycle is shown in a static position, and while it is detailed and well-lit, it doesn't seem to be in a state that suggests it is ready for use or has all its parts assembled. Wait, let's consider the context. The video is likely a promotional or showcase video for the motorcycle. Such videos often focus on the design and features of the motorcycle rather than showing it in a complete state. The absence of any indication of the motorcycle being ready for use or having all parts assembled suggests that it is not a complete motorcycle. <...> Final answer: B. ✗

Video-VER-7B: Let me think about this. The video showcases a sleek, modern motorcycle prominently displayed in a dimly lit garage. The motorcycle is brightly illuminated by a spotlight, highlighting its vibrant orange and black color scheme and intricate design features. Let's break down the question and the video content carefully. The question asks if the video shows a complete motorcycle. To answer this, we need to consider <...>. First, the video starts with a close-up of the motorcycle, focusing on its headlight and front design. The camera then pans around the motorcycle, showing different angles and details. This suggests that the video is indeed showing the motorcycle from various perspectives, which is a common practice when showcasing a complete vehicle. Next, the video includes a wider shot of the motorcycle in the garage, which further emphasizes the full view of the motorcycle. The garage setting and the presence of boxes in the background do not detract from the completeness of the motorcycle; they simply provide context for the setting. The video does not cut away to any other objects or scenes that would suggest a partial view of the motorcycle. Instead, it maintains a focus on the motorcycle throughout, highlighting its design and features in a comprehensive manner. Given this analysis, it is clear that the video does show a complete motorcycle. Answer is A. ✓

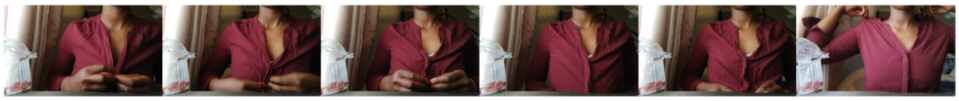
Question: How many times did the person launch objects on the table? A. 2 B. 3 C. 5 D. 4

Qwen2.5-VL-7B: Let me think about this. The person is seen launching objects on the table in the video. I can see that they have launched the object twice. The first time, they launch it and it bounces off the table. The second time, they launch it again and it bounces off the table as well. There are no other launches visible in the video. The answer is A. ✗

Video-R1-7B: Let me think about this carefully. The question is asking how many times the person launched objects on the table. To answer this, I need to recall the sequence of actions in the video. First, let's break it down: 1. The person launched an object, possibly a ball, onto the table. 2. The person then launched another object, possibly a different ball or a similar object, onto the table. Now, let's count the launches: - The first launch was clearly visible. - The second launch was also clearly visible. Hmm, it seems like there were two distinct launches. <...> Oh, I see. The person launched objects two times. The question is asking for <...>. Let's break it down one more time to be sure. The person launched an object, then launched another object. That's two launches. After this reflection, I am confident that the correct answer is two. Let's check the options: A. 2 B. 3 C. 5 D. 4 The correct answer is A. 2. ✗

Video-VER-7B: Let me think. The video shows a person interacting with objects on a wooden table. The person uses their hands to launch small objects across the table. The person launches the objects multiple times, and the objects roll across the table in different directions. The video captures the motion and trajectory of the objects as they are launched. Let's break down the sequence of events in the video to determine the number of times the person launched objects on the table. 1. The person first places a small object on the table and then uses their hand to launch it. This is the first launch. 2. The person then places another small object on the table and launches it. This is the second launch. 3. The person places a third small object on the table and launches it. This is the third launch. After reviewing the video, it is clear that the person launched objects three times. The person did not launch any objects more than three times, and there were no additional launches beyond these three. Now, let's consider the options: A. 2 - This is incorrect because the person launched objects more than twice. B. 3 - This is correct because the person launched objects three times. C. 5 - This is incorrect because the person did not launch objects five times. D. 4 - This is incorrect because the person did not launch objects four times. The correct answer is B, as the person launched objects three times. ✓

Figure 14: Comparison of reasoning traces from baseline models and our Video-VER model reveals that baselines often introduce speculative or hallucinated details not grounded in the video. In contrast, Video-VER consistently maintains alignment between reasoning steps and evidence. Text highlighted in red indicates incorrect reasoning steps, while text highlighted in green marks the key logical steps that lead to the correct answer.



Question: How many buttons did you see the person successfully buttoning up and correctly aligning the buttons with the holes? A. 2 B. 3 C. 5 D. 4

Qwen2.5-VL-7B: Let me think about this. The person is seen buttoning up a shirt, and it appears that they are successfully aligning the buttons with the holes. However, the video only shows a few moments of the buttoning process. It's not clear how many buttons were actually buttoned up and correctly aligned. The options provided are 2, 3, 5, and 4. Without seeing the entire process, it's difficult to determine the exact number. Hmm, the video only shows a few moments, so it's hard to be certain. Let's break it down: if we see 2 buttons being buttoned up, that would be option A. If we see 3, that would be B. If we see 4, that would be D. Option C, 5, seems unlikely based on the limited footage. Answer is A.

Video-R1-7B: Let me think about this carefully. The question is asking about the number of buttons the person successfully buttoned up and correctly aligned with the holes. I need to recall the details of the video and count the buttons. First, let's break it down. The person was seen buttoning up a shirt, and I need to count the buttons that were successfully aligned. From the video, I remember the person buttoning up the shirt in a specific sequence. Wait, let me verify. The person started from the bottom and worked their way up. I counted the buttons one by one. The first button was successfully aligned, the second button was also aligned, and the third button was aligned as well. Hmm, I need to double-check. The fourth button was also aligned, and the fifth button was aligned too. Oh, I see, the person successfully buttoned up five buttons in total. <...>. That's five buttons in total. After this reflection, I am confident that the person successfully buttoned up five buttons. The correct answer should be the option that corresponds to five buttons. Let's verify one last time. The person did not button up any buttons incorrectly, and all five buttons were aligned correctly. Therefore, the answer is C.

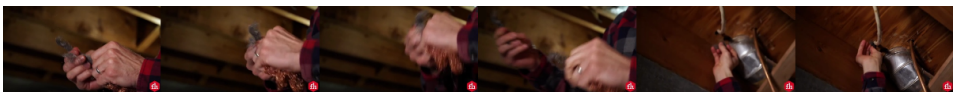
Video-VER-7B: Let me think. The video shows a person seated at a table, wearing a maroon blouse with a V-neckline. The person is seen buttoning up the blouse, carefully aligning each button with its corresponding hole. The process is methodical, with the person using both hands to ensure the buttons are properly aligned and secured. The background includes a curtain and a plastic bag on the table, adding context to the setting. The person successfully buttons up the blouse, completing the task with precision. Let's break down the video step by step to determine the number of buttons the person successfully buttoned up and correctly aligned with the holes. 1. The person starts by holding the blouse open, revealing the buttons. 2. The person then begins to button up the blouse, one button at a time. 3. The person carefully aligns each button with its corresponding hole and presses it down. 4. The person continues this process until the blouse is fully buttoned up. From the video, it appears that the person successfully buttoned up the blouse, and the number of buttons visible in the video is 4. The person aligned each button correctly with its hole and pressed it down, ensuring that the blouse was fully buttoned up. After careful consideration, the correct answer is D.

Question: Which description is a more suitable match for the video? A. The woman closest to the camera is moving at the fastest speed.
B. The woman closest to the camera is moving at the slowest speed.

Qwen2.5-VL-7B: The video shows a group of women running on a track. The woman closest to the camera appears to be running at a steady pace, but there's no indication that she is moving at the fastest speed. The other women in the group are also running at a similar pace, and there's no clear evidence that the woman closest to the camera is moving at the slowest speed. Therefore, the most suitable match for the video would be option B.

Video-R1-7B: Let me think about this for a moment. The question is asking <...>. First, let's break it down and consider the visual cues in the video. The woman closest to the camera appears to be running at a consistent pace, but there's no clear indication that she is moving at the fastest speed. In fact, the other runners seem to be maintaining a similar speed, and there's no evidence that the woman closest to the camera is lagging behind or speeding up. Now, let's consider the second option, <...>. Hmm, it seems that the woman closest to the camera is not moving at the fastest speed; she is moving at a similar pace to the other runners, which aligns with the second option. Final answer: B

Video-VER-7B: Let me think about this. The video showcases a group of female athletes participating in a track and field event, specifically a hurdles race. The runners are seen in various stages of their stride, with their bodies leaning forward and their arms swinging rhythmically. <...> The runners are moving at a consistent pace, with some slightly ahead of others, suggesting a competitive race. The question asks <...>. First, let's consider the spatial details. The woman closest to the camera <...> appears to be slightly ahead of the others. <...> Next, let's consider the temporal details. The video captures the runners in various stages of their stride, and the woman closest to the camera is not lagging behind <...> is maintaining a steady pace, which is typical for a competitive race <...>. Therefore, the more suitable match for the video is that the woman closest to the camera is moving at a faster speed. Final answer: A



Question: Is the person in the video holding metal wires? A. yes B. no

Qwen2.5-VL-7B: Let me think about this. The video shows a person's hands working with a tool, and there are sparks flying, which suggests some kind of metalworking activity. However, the video does not clearly show <...>. The hands are manipulating something, but it's not explicitly <...>. The person is also seen handling a metal container, which could be related to the activity but doesn't necessarily indicate <...>. The video does not provide clear evidence of the person holding metal wires. The activity seems to involve metalworking, but the specific object being manipulated is not identified as metal wires.

Video-R1-7B: Let me think about this for a moment. The question is asking <...>. To answer this, I need to carefully analyze the visual information provided in the video. First, let's break it down. The video shows a person working with some kind of tool, and there are sparks flying around. This suggests that the person is likely using a tool that involves cutting or grinding metal. The person is also seen holding a cylindrical object, which appears to be a metal container or a part of the tool. Now, let's consider the wires. The wires mentioned in the question are not explicitly shown in the video. The person is not holding any wires that are clearly visible. The focus is on the tool and the sparks, which are indicative of metalworking. Hmm, it seems that the person is not holding metal wires. The wires are not a prominent feature in the video, and the person is not manipulating any wires. The person is working with a tool that involves metal, but the wires are not the focus of the video. <...>, I am confident that the person is not holding metal wires. <...> Final answer: B

Video-VER-7B: Let me think. <...> The video clearly shows a person working with wires, and the question asks if the person <...>. The person is indeed handling wires, and the wires appear to be metallic based on their sheen and the way they are being manipulated. The person's hands are carefully gripping the wires, and there is no indication that the wires are not metallic. The workshop setting further supports the idea <...>. Now, let's consider the options. Option A suggests that the person is holding metal wires, which aligns with the visual evidence in the video. Option B suggests that the person is not holding metal wires, which contradicts the visual evidence. The person's actions and the environment strongly support the idea that the wires are metallic. <...> Given the detailed analysis of the video and the question, the answer is A.

Figure 15: Comparison of reasoning traces from baseline models and our Video-VER model reveals that baselines often introduce speculative or hallucinated details not grounded in the video. In contrast, Video-VER consistently maintains alignment between reasoning steps and evidence. Text highlighted in red indicates incorrect reasoning steps, while text highlighted in green marks the key logical steps that lead to the correct answer.