VIDEO-THINKER: SPARKING "THINKING WITH VIDEOS" VIA REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in multimodal reasoning, particularly Thinking with Images, have achieved remarkable progress by empowering Multimodal Large Language Models (MLLMs) to actively inspect and reason over visual context. Extending this visual reasoning paradigm to dynamic videos requires models to actively localize and synthesize visual evidence across the temporal dimension. However, enabling MLLMs to think with videos faces two fundamental challenges: (i) from the data perspective, existing training data lacks tasks that guide models to actively seek and verify evidence across temporal intervals; and (ii) from the training supervision perspective, current methods lack verified temporal process supervision, leading to unfaithful reasoning and superficial shortcuts. To address these challenges, we propose Video-Thinker to unify data construction and training supervision, enabling intrinsic temporal reasoning without relying on external tools. From the data perspective, to formally capture both temporal and semantic dependencies across events, we construct a video reasoning graph to model their temporal ordering and semantic relations. Leveraging this graph, we synthesize reasoning tasks with guaranteed evidence necessity, ensuring that correct answers strictly require observing and integrating all designated temporal clues. From the training supervision perspective, following a cold-start Supervised Fine-Tuning (SFT) phase, we introduce an evidence-aware Reinforcement Learning (RL) method that provides verified temporal process supervision to ensure faithful grounding throughout reasoning. Extensive experiments across six challenging video reasoning benchmarks demonstrate that Video-Thinker achieves an average improvement of +6.86% over the base model with remarkable data efficiency, requiring only 20K training instances.

1 INTRODUCTION

Multimodal Large Language Models have witnessed a profound paradigm shift toward Thinking with Images, transitioning vision from a passive, static input into an active cognitive tool (OpenAI, 2024; Zheng et al., 2025; Shen et al., 2024; Wang et al., 2025d; Ma et al., 2024). By dynamically inspecting, localizing, and interleaving visual observations within reasoning traces, models have established unprecedented performance across visual question answering (Liu et al., 2023; Zhao et al., 2025; Gupta & Kembhavi, 2023), mathematical problem solving (Chen et al., 2025c; Shao et al., 2024a; Wang et al., 2025a; Yue et al., 2024), and complex scene understanding (You et al., 2023; Yang et al., 2023; Zhang et al., 2025b; Ma et al., 2025). The natural frontier of this cognitive paradigm lies in video understanding. Fundamentally, video reasoning is an active process of seeking, localizing, and synthesizing temporal evidence across dynamic video events (Wang & Peng, 2025; Dang et al., 2025; Yu et al., 2025a). While recent efforts often rely on external tools or auxiliary modules to navigate video content (Liu et al., 2025a), such pipelined architectures suffer from cascading error propagation and contextual disconnection across isolated components. Consequently, a promising yet challenging direction is to cultivate an intrinsic capability of Thinking with Videos, empowering models to autonomously localize temporal intervals, extract visual clues, and synthesize evidence directly within their native chain-of-thought processes.

Despite recent efforts to post-train video reasoning models via reinforcement learning (Feng et al., 2025; Li et al., 2025b;a), existing approaches suffer from severe data inefficiency, routinely requiring massive corpora exceeding 165K samples (Feng et al., 2025) while remaining vulnerable to temporal hallucinations. We argue that this deficiency is rooted in two fundamental challenges across data and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

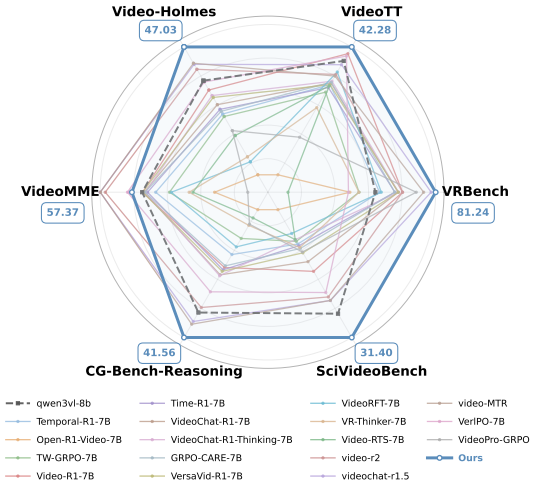

Figure 1: Performance comparison of VIDEO-THINKER across six challenging video reasoning benchmarks against the base model and representative post-trained reasoning models.

supervision: First, from the data perspective, existing training data lacks tasks that guide models to actively seek and verify evidence across temporal intervals. Conventional data synthesis pipelines query language models over flat sequences of event captions without structural constraints (Fei et al., 2024). Lacking an explicit structural topology, flat text cannot formalize informational subsumption. Consequently, multi-segment questions frequently suffer from evidence redundancy, allowing models to deduce answers through text commonsense alone or by inspecting an isolated clip, without genuinely perceiving the required temporal events. Second, from the training supervision perspective, prevailing methods optimize policies solely using coarse outcome rewards on final answers (Shao et al., 2024b). When training relies strictly on final correctness, models can easily receive positive reinforcement for lucky guesses despite generating entirely incorrect or hallucinated time intervals. This reinforces unfaithful reasoning and superficial shortcuts, encouraging models to mimic formatting rather than perform faithful temporal grounding.

To address these challenges, we propose VIDEO-THINKER to unify data construction and training supervision, enabling intrinsic temporal reasoning without relying on external tools. From the data perspective, to formally capture both temporal and semantic dependencies across events, we introduce the video reasoning graph to model their temporal ordering via Allen's interval algebra (Allen, 1983)and semantic dependencies via propositional entailment. By merging mutually entailing events into canonical clusters, the graph provides an unambiguous topological foundation. Leveraging this graph, we synthesize reasoning tasks with guaranteed evidence necessity, where alternative conclusions are derived from partial evidence subsets, ensuring that correct answers strictly require observing and integrating all designated temporal clues.

From the training supervision perspective, following a cold-start Supervised Fine-Tuning (SFT) phase over 10K structured reasoning trajectories interleaving temporal localization (`<time>`) and visual description (`<caption>`), as shown in Appendix H (Figure 5), we advance the model via an evidence-aware Reinforcement Learning (RL) method over 10K task instances. Rather than adding heuristic grounding rewards that alter the optimal policy, we formulate an evidence-aligned multi-span grounding score using temporal Intersection-over-Union against certified evidence spans, and inject it as a mean-centered perturbation directly into the advantage function. We analyze the boundary condition on the evidence weight to ensure that process feedback sharpens credit assignment without penalizing factually correct answers, thereby suppressing unfaithful reasoning and cultivating faithful temporal grounding.

Extensive experiments across six challenging video reasoning benchmarks—Video-Holmes (Cheng et al., 2025), CG-Bench-Reasoning (Chen et al., 2024), VRBench (Yu et al., 2025b), SciVideoBench (Deng et al., 2025), VideoTT (Zhang et al., 2025a), and VideoMME (Fu et al., 2024)—demonstrate that VIDEO-THINKER, built upon Qwen3-VL-8B-Instruct (Bai et al., 2026), achieves substantial performance gains across diverse video reasoning tasks. It attains an average score of 50.15% across all six benchmarks, delivering a decisive +6.86% absolute improvement over the base model and outperforming previous post-trained reasoning models (Figure 1). Crucially,

VIDEO-THINKER achieves these gains using only 20K training instances, split evenly between 10K for SFT cold-start and 10K for process-supervised RL, demonstrating that enforcing structural evidence necessity on video reasoning graphs and providing process-level credit assignment yields far superior data efficiency and reasoning fidelity.

2 RELATED WORK

Multimodal Large Language Models are shifting from treating vision as passive context toward using visual operations as active cognitive tools (OpenAI, 2024; Zheng et al., 2025; Liu et al., 2025b), where models dynamically inspect and interleave visual observations within reasoning traces (Jaech et al., 2024; Guo et al., 2025; Su et al., 2025). Extending this paradigm to video, recent efforts apply reinforcement learning to cultivate temporal reasoning (Feng et al., 2025; Li et al., 2025b;a; Wang & Peng, 2025; Dang et al., 2025; Yu et al., 2025a). However, existing methods predominantly rely on coarse outcome rewards over final answers (Shao et al., 2024b), which inadvertently reinforce superficial shortcuts and hallucinated timestamps, while their training data lacks structural guarantees that questions strictly necessitate temporal evidence. Meanwhile, tool-augmented approaches that deploy external modules for video navigation (Liu et al., 2025a) suffer from cascading errors and contextual disconnection. In contrast, VIDEO-THINKER advances an endogenous paradigm by internalizing temporal localization and visual description directly into the model's native chain of thought, supported by certified task synthesis and process-level supervision. An extended discussion of related work is provided in Appendix A.

3 THE VIDEO-THINKER FRAMEWORK

We consider an untrimmed video sequence V accompanied by a collection of localized event segments sourced from four dense video corpora (MDVC, VideoXum, VidSTG, and YouCook2, detailed in Appendix C), where each segment couples a continuous temporal interval $I = [t^s, t^e]$ with a descriptive textual caption c. Note that these segment annotations are exclusively utilized during offline graph construction for data synthesis; during inference, VIDEO-THINKER operates in a fully autonomous, end-to-end manner on raw untrimmed videos without requiring any external segment priors. While conventional synthesis pipelines directly prompt language models over unstructured segment lists (Fei et al., 2024; Cheng et al., 2025), they struggle to guarantee that downstream reasoning strictly necessitates designated temporal evidence. To address this limitation, VIDEO-THINKER transforms disjoint event segments into a structured video reasoning graph (§3.1), synthesizes reasoning tasks with certified necessary evidence (§3.2), and trains the Multimodal Large Language Model (MLLM) via evidence-aware process supervision (§3.3), as summarized in Algorithm 1 (with complete pipeline pseudo-code in Appendix B).

3.1 VIDEO REASONING GRAPH AS AN EVENT REPRESENTATION

To capture both qualitative temporal progression and semantic information containment across video segments, we formalize the video into an explicit topological representation:

Definition 3.1 (Video Reasoning Graph)**.** *Given an untrimmed video V associated with n event segments, we represent V as a directed attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_{temp}, \mathcal{E}_{ent})$:*

- *$\mathcal{V} = \{v_i\}_{i=1}^n$ is the set of video event nodes, where each node $v_i = (I_i, c_i)$ couples a continuous temporal interval $I_i = [t_i^s, t_i^e]$ ($t_i^s < t_i^e$) with a descriptive textual caption c_i;*
- *$\mathcal{E}_{temp}$ denotes pairwise qualitative temporal relations between event intervals governed deterministically by Allen's interval algebra; and*
- *$\mathcal{E}_{ent}$ denotes directed propositional entailment edges, where directed edge $(u \Rightarrow v) \in \mathcal{E}_{ent}$ asserts that the factual occurrence of event u logically guarantees the realization of event v.*

Deterministic Temporal Relations via Allen's Algebra. For any event pair $(u, v) \in \mathcal{V} \times \mathcal{V}$, their qualitative temporal alignment is computed deterministically via Allen's interval algebra (Allen, 1983). Allen's algebra defines 13 mutually exclusive and exhaustive qualitative relations between

two temporal intervals:

$$\mathcal{R}_{\text{Allen}} = \big\{\textit{before}, \textit{after}, \textit{meets}, \textit{met-by}, \textit{overlaps}, \textit{overlapped-by}, \textit{during}, \textit{contains}, \textit{starts}, \textit{started-by}, \textit{finishes}, \textit{finished-by}, \textit{equals}\big\}. \tag{1}$$

Because this step relies exclusively on scalar comparisons between interval endpoints $[t_u^s, t_u^e]$ and $[t_v^s, t_v^e]$, $\mathcal{E}_{\text{temp}}$ is computed with $O(|\mathcal{V}|^2)$ complexity without invoking neural models, completely averting temporal hallucinations. Crucially, the induced temporal subgraphs dictate the syntactic form of downstream reasoning questions: sequential relations (*before*, *after*) support causal and order queries, concurrent relations (*overlaps*, *overlapped-by*) support synchronization queries, and containment relations (*during*, *contains*) support background-conditioned queries. Question diversity thus emerges organically from graph topology rather than arbitrary hand-crafted taxonomies.

Propositional Entailment and Event Canonicalization. While temporal edges govern dynamic progression, semantic redundancy occurs when one event's description logically subsumes another. For all candidate event pairs within a video, we query Gemini-3.1-Flash (DeepMind, 2026) in a single unified call to evaluate directional propositional entailment ($u \Rightarrow v$, directed from specific to general) using an atomic factual checklist protocol that verifies whether all atomic assertions of v are strictly entailed by u within a single inferential step (see Appendix G.2 for the complete prompt and labelling protocol). Evaluating all candidate pairs simultaneously avoids quadratic round-trip overhead while ensuring a globally consistent relation topology. When two descriptions mutually entail each other ($u \Rightarrow v \wedge v \Rightarrow u$), they represent identical physical occurrences articulated with alternative phrasing. We merge such mutually entailing events into a canonical event cluster, where the cluster retains the union of their physical intervals. This canonicalization eliminates semantic cycles, yielding a cycle-free canonical event set $\hat{\mathcal{V}}$ where entailment edges form a directed acyclic graph (DAG).

3.2 Training Data Synthesis over Video Reasoning Graphs

Given the canonical event set $\hat{\mathcal{V}}$ and temporal relations $\mathcal{E}_{\text{temp}}$, Video-Thinker synthesizes certified training instances comprising contrastive questions, ground-truth evidence spans, and full reasoning trajectories.

Non-Redundant Evidence Sampling. To construct a reasoning task requiring k distinct visual cues, we sample a candidate evidence subset $A = \{s_1, \ldots, s_k\} \subset \hat{\mathcal{V}}$. For multi-segment questions ($k \geq 2$), to ensure that every selected segment is indispensable, we enforce that all events in A are pairwise non-entailing: $\forall u \neq v \in A \implies \neg(u \Rightarrow v) \wedge \neg(v \Rightarrow u)$. This condition structurally eliminates information redundancy: if $u \Rightarrow v$, observing segment u already provides the information in v, making segment v superfluous. Any candidate subset violating this condition is directly filtered out on the graph before generating questions. For single-cue questions ($k = 1$), candidate events are sampled directly from $\hat{\mathcal{V}}$ with non-entailing alternative events within the same video used as distractors.

QA Generation with Necessary Evidence. Given a non-redundant evidence set A and its induced temporal signature $\sigma = \{\tau(u, v) \in \mathcal{R}_{\text{Allen}} \mid u, v \in A\}$ specifying pairwise Allen relations between evidence segments, we formulate question Q and candidate options. Let $\mathcal{R}_Q(W)$ denote the most specific factual conclusion an observer can establish for question Q when observing only segment subset $W \subseteq A$. We formulate the ground-truth answer y and contrastive distractors $\{d_i\}_{i=1}^k$ via slot omission:

$$y = \mathcal{R}_Q(A), \qquad d_i = \mathcal{R}_Q\big(A \setminus \{s_i\}\big), \quad \forall s_i \in A. \tag{2}$$

Intuitively, correct answer y represents the comprehensive synthesis requiring full evidence set A, whereas each distractor d_i is precisely the valid deduction reachable when slot s_i is omitted and only the remaining $k-1$ segments are witnessed. For instance, in a sequential procedure where an ingredient is mixed (s_1) before baking (s_2), omitting s_1 induces the flawed distractor that baking occurred directly without prior preparation. Whenever $d_i \neq y$, any agent failing to observe segment s_i lacks the empirical evidence to distinguish y from d_i, bounding its decision between that pair to random guessing. To prevent reliance on caption completeness, we prioritize structural distractors that permute verified temporal orderings and agent bindings certified by timestamps. Candidate tasks are accepted only when every distractor is factually distinct from the ground truth ($d_i \neq y$) and passes blind-attack filtering (discarding items correctly answerable above chance without video input). Operationally, rather than prompting the generator independently for each slot, we query

Algorithm 1 VIDEO-THINKER Training Pipeline

Require: Video V, segments $\{v_i\}$, policy π_θ, group size G, weight $\lambda \leq 1/\sqrt{G-1}$.
Ensure: Optimized policy π_θ.
1: Construct video reasoning graph $\mathcal{G} = (\hat{\mathcal{V}}, \mathcal{E}_{\text{temp}}, \mathcal{E}_{\text{ent}})$ via Allen algebra and entailment (§3.1)
2: Sample pairwise non-entailing evidence $A \subset \hat{\mathcal{V}}$ and synthesize contrastive tasks with certified spans $\mathcal{E}^*$ (§3.2)
3: Cold-start π_θ on structured trajectories $\mathcal{D}_{\text{SFT}}$ by minimizing SFT loss $\mathcal{L}_{\text{SFT}}(\theta)$ (Eq. 4)
4: **for** each GRPO iteration on $(V, Q, y, \mathcal{E}^*)$ **do**
5: Sample group of G reasoning outputs $\{T^{(i)}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(T \mid V, Q)$
6: Compute multi-span evidence score $e^{(i)} = \min_j \max_{I \in \mathcal{I}(s_j)} \max_{c \in C^{(i)}} \text{tIoU}(c, I)$ (Eq. 6)
7: Formulate advantage $A_i = z(a^{(i)}) + \lambda(e^{(i)} - \bar{e})$ via credit redistribution (Eq. 7)
8: Update policy parameter θ by maximizing GRPO surrogate objective $\mathcal{J}_{\text{GRPO}}(\theta)$ (Eq. 5)
9: **end for**
10: **return** Optimized policy π_θ.

Gemini-3.1-Flash with a unified prompt containing evidence set A and its induced temporal signature σ, encouraging distractors to be mutually distinct and balanced in length. Crucially, question text Q strictly omits explicit numerical timestamps or trivial index pointers (e.g., "the first event"), ensuring that questions cannot be resolved through text matching alone without actively inspecting visual video frames.

Structured Trajectory and Certified Evidence Spans. Beyond question-answer pairs, VIDEO-THINKER generates complete reasoning trajectories to bootstrap model training. Conditioned on video V, question Q, ground-truth answer y, and evidence set A, we prompt Gemini-3.1-Flash (DeepMind, 2026) to synthesize an interleaved chain-of-thought trajectory T. As illustrated in Appendix H (Figure 5), each key segment is systematically articulated through two specialized tags: the `<time></time>` tag denoting temporal localization and the `<caption></caption>` tag articulating observed visual cues. Concurrently, each synthesized task registers its certified evidence intervals $\mathcal{E}^* = \{\mathcal{I}(s_1), \ldots, \mathcal{I}(s_k)\}$, where:

$$\mathcal{I}(s_j) = \bigcup_{u \in \hat{\mathcal{V}},\, u \Rightarrow s_j} \mathcal{T}(u). \tag{3}$$

Here, $\mathcal{T}(u)$ denotes the continuous physical temporal interval(s) associated with canonical event u, and $\mathcal{I}(s_j)$ collects all intervals from events that entail s_j. This captures evidence equivalence: if a model localizes an alternative occurrence u that logically entails s_j (e.g., localizing "a man holds a cup filled with water" when the target slot states "a man holds a cup"), witnessing u conclusively validates s_j, providing faithful supervision without penalizing alternative valid grounding.

3.3 TRAINING WITH EVIDENCE-AWARE PROCESS SUPERVISION

Equipped with curated reasoning trajectories and certified evidence intervals, VIDEO-THINKER trains the MLLM through a two-stage training paradigm, as detailed in Algorithm 1.

Stage 1: Cold-Start Supervised Fine-Tuning. Pre-trained MLLMs inherently lack inductive bias for structured temporal navigation using specialized tags (`<time>` and `<caption>`). We therefore establish a cold-start initialization via Supervised Fine-Tuning (SFT) over curated trajectories $\mathcal{D}_{\text{SFT}} = \{(V, Q, T, y)\}$. The SFT objective minimizes the negative log-likelihood of generating reasoning trajectory T and final answer y:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(V,Q,T,y)\sim\mathcal{D}_{\text{SFT}}} \sum_{t=1}^{|[T;y]|} \log p_\theta\big([T;y]_t \mid V, Q, [T;y]_{<t}\big), \tag{4}$$

where $[T;y]$ denotes token concatenation, and p_θ is the MLLM policy parameterized by θ. SFT aligns the model with the interleaved formatting syntax and basic spatiotemporal reasoning priors.

Stage 2: Process-Supervised Policy Optimization via GRPO. While SFT teaches formatting adherence, it suffers from exposure bias and cannot autonomously optimize dynamic temporal search strategies. We advance the model via Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) augmented with certified process supervision. For each task $(V, Q, y, \mathcal{E}^*)$, the policy samples a group of G reasoning outputs $\{T^{(1)}, \ldots, T^{(G)}\}$ from the previous policy $\pi_{\theta_{\text{old}}}$. The policy is

optimized by maximizing the clipped surrogate objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(V,Q,y,\mathcal{E}^*)} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\min \left(\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}} A_i,\ \text{clip} \left(\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}, 1-\epsilon, 1+\epsilon \right) A_i \right) - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right) \right], \tag{5}$$

where $\pi_\theta = p_\theta(T^{(i)} \mid V, Q)$, π_{ref} is the reference policy, and A_i is the process-supervised advantage. For completeness, each trajectory receives standard outcome-format feedback $r^{(i)} = a^{(i)} + r^{(i)}_{\text{format}}$, where $a^{(i)} \in \{0, 1\}$ indicates answer correctness and $r^{(i)}_{\text{format}}$ penalizes formatting violations.

Instead of relying purely on binary outcome rewards or adding heuristic grounding penalties into rewards (which alters the optimal policy and encourages reward hacking), we formulate an evidence-aligned multi-span grounding score $e^{(i)} \in [0, 1]$ and inject it directly into the advantage function. Let $C^{(i)} = \{c_1, \ldots, c_m\}$ denote the temporal intervals parsed from the `<time>` tags of trajectory $T^{(i)}$. Given certified evidence spans $\mathcal{E}^* = \{\mathcal{I}(s_1), \ldots, \mathcal{I}(s_k)\}$, the evidence score is computed as:

$$e^{(i)} = \min_{1 \le j \le k} \max_{I \in \mathcal{I}(s_j)} \max_{c \in C^{(i)}} \text{tIoU}(c, I), \tag{6}$$

where $\text{tIoU}(c, I) = \frac{|c \cap I|}{|c \cup I|}$ is the temporal Intersection-over-Union, with $e^{(i)} = 0$ if $C^{(i)} = \emptyset$. Equation (6) cleanly unifies three foundational evidence topologies without requiring ad-hoc heuristic branching: (i) **All-required evidence (conjunction):** for multi-slot tasks ($k \ge 2$) where reasoning strictly necessitates observing all distinct intervals (e.g., sequential or concurrent events), the outer $\min_{1 \le j \le k}$ enforces complete coverage—omitting even a single critical slot collapses the overall evidence score to zero; (ii) **Either-suffices evidence (disjunction):** when an event is witnessed across multiple equivalent intervals in $\mathcal{I}(s_j)$ (e.g., coarse versus fine-grained descriptions of the same action), the middle $\max_{I \in \mathcal{I}(s_j)}$ accepts grounding against any valid certified instance, ensuring faithful supervision without penalizing alternative valid observations; and (iii) **Single-interval evidence:** for localized single-cue queries ($k = 1$), the formulation naturally reduces to standard single-span tIoU. Importantly, trivial shortcuts such as predicting a span across the entire video $[0, T]$ naturally collapse the score through the union denominator $|c \cup I|$, precluding reward hacking without manual penalty tuning.

The advantage A_i is then constructed via group credit redistribution:

$$A_i = \underbrace{z\big(a^{(i)}\big)}_{\text{Outcome Standardization}} + \lambda \underbrace{\big(e^{(i)} - \bar{e}\big)}_{\text{Mean-Centered Process Evidence}}, \tag{7}$$

where $z(a^{(i)}) = \frac{a^{(i)} - \text{mean}(\{a^{(j)}\})}{\text{std}(\{a^{(j)}\}) + \delta}$ is group-standardized outcome feedback, and $\bar{e} = \frac{1}{G} \sum_{j=1}^{G} e^{(j)}$ is the group mean evidence score. Crucially, because $\sum_{i=1}^{G} (e^{(i)} - \bar{e}) = 0$, process evidence acts as a zero-sum credit redistribution across trajectories sharing identical outcome correctness, preserving the global optimality of task accuracy while favoring well-grounded reasoning paths. In edge cases where all sampled trajectories within a group share identical correctness ($\text{std}(a) = 0$), if all responses are correct, process evidence naturally acts as a tie-breaker to reinforce well-grounded reasoning; if all responses are incorrect, the process term is muted ($\lambda = 0$) to prevent rewarding hallucinated reasoning traces that yield wrong answers.

To ensure that process supervision does not penalize factually correct answers, we derive a boundary condition on λ (detailed in Appendix I), establishing that:

$$\lambda \le \frac{1}{\sqrt{G-1}} \tag{8}$$

guarantees $A_i > 0$ for all correct trajectories, keeping process feedback aligned with answer correctness.

Table 1: Comparison of model performance across six video reasoning benchmarks. # Data denotes the scale of post-training data. ✔ and ✘ denote the presence and absence of temporal supervision during training, respectively. The best results are marked in **red bold** and second-best in blue underline.

Model	Size	Temp. Sup.	# Data	Video-Holmes	CG-Bench	VRBench	SciVideo	VideoTT	VideoMME	Avg.
Base Model										
Qwen3-VL-Instruct	8B	–	–	40.39	36.66	59.40	28.70	39.70	54.89	43.29
Post-trained Reasoning Models										
Temporal-R1	7B	✔	32K	33.81	25.27	60.92	20.80	35.31	51.63	37.96
Open-R1-Video	7B	✘	4K	21.83	16.46	50.15	16.90	18.66	37.41	26.90
TW-GRPO	7B	✘	1K	33.32	22.11	53.46	20.50	33.93	48.30	35.27
Video-R1	7B	✔	165K	38.54	27.81	69.25	23.90	41.04	54.81	42.56
Time-R1	7B	✔	2.5K	34.73	28.28	66.48	21.00	34.78	53.59	39.81
VideoChat-R1	7B	✔	18K	35.65	29.26	67.65	22.80	35.63	54.41	40.90
VideoChat-R1-Thinking	7B	✔	18K	37.45	29.44	67.81	20.30	35.95	54.15	40.85
GRPO-CARE	7B	✘	6K	34.34	27.49	66.39	21.30	35.74	54.22	39.91
VersaVid-R1	7B	✘	10.8K	37.07	28.58	67.72	21.80	35.31	54.78	40.88
VideoRFT	7B	✘	310K	24.39	23.77	61.54	19.60	37.65	47.85	35.80
VR-Thinker	7B	✘	20K	25.37	19.54	53.43	21.20	31.07	43.56	32.36
Video-RTS	7B	✘	6K	29.56	18.09	27.71	20.30	35.10	42.63	28.90
Video-R2	7B	✔	15.3K	42.62	35.67	69.22	26.80	37.20	63.80	45.88
VideoChat-R1.5	7B	✔	80K	43.60	38.42	79.61	27.20	39.00	**65.04**	48.81
Video-MTR	7B	✔	8K	43.82	38.95	76.90	27.20	36.90	65.00	48.13
VerlPO	7B	✘	117K	40.12	32.59	49.62	26.30	40.60	58.44	41.28
VideoPro-GRPO	8B	✘	13.5K	30.54	19.31	74.20	21.70	25.60	29.41	33.46
Our Framework										
Video-Thinker (Ours)	8B	✔	20K	**47.03**	**41.56**	**81.24**	**31.40**	**42.28**	57.37	**50.15**

Table 2: Ablation study on training components across six video reasoning benchmarks. ✔ and ✘ denote the presence and absence of temporal supervision during training, respectively. The best results are marked in **red bold** and second-best in blue underline.

Configuration	Temp. Sup.	Video-Holmes	CG-Bench	VRBench	SciVideo	VideoTT	VideoMME	Avg.
Base Model								
Qwen3-VL-Instruct	–	40.39	36.66	59.40	28.70	39.70	54.89	43.29
Ablation Configurations								
+ SFT	✘	38.74	31.83	61.19	25.60	37.54	52.19	41.18
+ Pure RL	✘	37.41	31.25	62.38	24.90	38.16	50.74	40.81
+ SFT + Outcome-based RL	✘	45.28	37.25	78.34	28.40	**42.88**	55.32	47.91
Full Framework								
+ SFT + Process-supervised RL (Ours)	✔	**47.03**	**41.56**	**81.24**	**31.40**	42.28	**57.37**	**50.15**

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Benchmarks and Evaluation Protocol. To comprehensively assess the video reasoning capabilities of VIDEO-THINKER, we conduct extensive evaluations across six diverse, challenging video reasoning benchmarks: Video-Holmes (Cheng et al., 2025), CG-Bench-Reasoning (Chen et al., 2024), VRBench (Yu et al., 2025b), SciVideoBench (Deng et al., 2025), VideoTT (Zhang et al., 2025a), and VideoMME (Fu et al., 2024). These benchmarks probe distinct cognitive dimensions, ranging from long-horizon multi-clue detective deduction and fine-grained procedural event tracking to scientific experiment interpretation and narrative short-video comprehension. Details of all benchmark datasets are provided in Appendix D. Across all evaluations, we maintain a strictly standardized decoding protocol: temperature 0.1, top-p 0.001, and identical prompt templates (see Appendix G). To guarantee a strictly fair comparison, all baseline models and VIDEO-THINKER are evaluated under this identical inference protocol, uniformly subsampling each video to a maximum budget of 16 frames during inference.

Baseline Models. We evaluate VIDEO-THINKER against the pre-trained Qwen3-VL-8B-Instruct base model (Bai et al., 2026) and 17 representative post-trained video reasoning models spanning diverse optimization paradigms: Temporal-R1 (Li et al., 2025a), Open-R1-Video (Wang & Peng, 2025), TW-GRPO (Dang et al., 2025), Video-R1 (Feng et al., 2025), Time-R1 (Wang et al., 2025c), VideoChat-R1 and VideoChat-R1-Thinking (Li et al., 2025b), GRPO-CARE (Chen et al., 2025d),

VersaVid-R1 (Chen et al., 2025b), VideoRFT (Wang et al., 2026), VR-Thinker (Wang et al., 2025b), Video-RTS (Wang et al., 2025e), Video-R2 (Maaz et al., 2025), VideoChat-R1.5 (Yan et al., 2026), Video-MTR (Xie et al., 2025), VerlPO (Li et al., 2025c), and VideoPro-GRPO (Li et al., 2026) (summarized in Table 1). Detailed architectural characteristics and training configurations for all baselines are provided in Appendix E.

Training Details. We employ Qwen3-VL-8B-Instruct (Bai et al., 2026) as our foundational multimodal backbone. Our post-training dataset contains exactly 20K instances, partitioned into 10K structured reasoning trajectories for SFT cold-start and 10K task instances for process-supervised RL (detailed source dataset composition and segment distributions in Appendix C and Table 4). During the SFT cold-start stage, we fine-tune the model for 1 epoch with a learning rate of 1×10^{-5} and a batch size of 16 to instill the structural syntax of interleaving temporal localization (<time>) and visual description (<caption>). For the subsequent process-supervised RL stage, we optimize policy parameters via GRPO with candidate group size $G = 8$, clipping parameter $\epsilon = 0.2$, KL divergence penalty coefficient $\beta = 0.04$, weight decay of 0.01, and maximum gradient clipping norm of 5. The process evidence weight is set to $\lambda = 0.25$, which strictly respects the theoretical safety boundary $\lambda \leq 1/\sqrt{G-1} \approx 0.378$ established in Eq. 8. The initial policy learning rate is set to 5×10^{-6} with a batch size of 8. Both training stages utilize the unified prompt template detailed in Appendix G. During training, video inputs are sampled at FPS $= 1$ up to a maximum of 384 frames (with each frame dynamically encoded at a resolution of $128 \times 28 \times 28$ pixels) to provide the fine-grained temporal resolution required to supervise precise <time> interval boundaries. Evaluating under constrained 16 frames at test time thus rigorously benchmarks the policy's capacity for faithful temporal reasoning under temporally sparse observations. All post-training experiments are conducted on 8 NVIDIA H800 (80GB) GPUs.

4.2 Performance Comparisons and Analysis

We evaluate Video-Thinker alongside all baseline models across the six benchmarks using standard top-1 accuracy (%). The comprehensive evaluation results are summarized in Table 1, yielding the following key insights.

Video-Thinker achieves state-of-the-art performance with exceptional data efficiency. As reported in Table 1, Video-Thinker attains an overall average accuracy of 50.15% across the six benchmarks, delivering a decisive +6.86% absolute improvement over the pre-trained Qwen3-VL-Instruct base model (43.29%). Remarkably, Video-Thinker establishes the highest performance on five of the six benchmarks: Video-Holmes (47.03%), CG-Bench-Reasoning (41.56%), VRBench (81.24%), SciVideoBench (31.40%), and VideoTT (42.28%), while maintaining strong, competitive accuracy on VideoMME (57.37%).

Breaking down performance across domains reveals three primary strengths: First, on complex multi-clue deduction (Video-Holmes 47.03%, CG-Bench 41.56%), Video-Thinker outperforms the strongest baseline Video-MTR (43.82%, 38.95%) by +3.21% and +2.61%, validating that pairwise non-entailing task synthesis eliminates superficial language shortcuts. Second, on fine-grained temporal tracking (VRBench 81.24%, SciVideoBench 31.40%), our model sets new state-of-the-art results (+1.63% over VideoChat-R1.5 on VRBench and breaking the 30% threshold on scientific dynamics), confirming that mean-centered credit redistribution sharpens interval boundary sensitivity. Finally, on fast-paced shorts (VideoTT 42.28%) and broad multimodal QA (VideoMME 57.37%), Video-Thinker exhibits robust cross-domain generalization (+2.48% over the base model on VideoMME).

Crucially, these gains are achieved with unprecedented data efficiency. While competitive reasoning models require massive post-training corpora—such as VideoChat-R1.5 (80K) or Video-R1 (165K)—Video-Thinker establishes superior average accuracy using merely 20K training instances (10K SFT + 10K RL). With only 12.1% of Video-R1's data, Video-Thinker outperforms Video-R1 by +7.59% on average (50.15% vs. 42.56%), demonstrating that enforcing structural evidence necessity on video reasoning graphs enables far more effective reasoning transfer than uncurated data scaling.

Ablation study on training components. Table 2 isolates the contributions of SFT cold-start, outcome-based RL, and our proposed process-supervised RL across the six benchmarks: (i) *Isolated Training Fails*: SFT alone (41.18%) and pure RL (40.81%) both fall below the base model (43.29%). SFT incurs an alignment tax by over-constraining generation syntax without exploratory search,

Table 3: Comparison of video grounding and captioning performance on 1K samples from ActivityNet, YouCook2, and TutorialVQA. The best results are marked in **red bold** and second-best in blue underline.

Model	Grounding				Captioning			
	mIOU	Recall@0.3	Recall@0.5	Average	Meteor	ROUGE-L	BLEU@1	Average
Qwen3-VL-8B-Instruct	32.45	45.82	29.14	35.80	15.18	16.75	12.36	14.76
VideoChat-R1.5	41.68	70.45	43.27	51.80	16.14	19.28	14.52	16.65
Video-Thinker (Ours)	**52.74**	**84.62**	**58.35**	**65.24**	**17.85**	**22.94**	**17.91**	**19.57**

while pure RL suffers severe reward sparsity and format collapse over long video sequences from scratch. (ii) *Two-Stage Synergy*: SFT cold-start combined with outcome-based RL (+SFT + Outcome-based RL, 47.91%) unlocks substantial reasoning capabilities (+6.73% over SFT), proving that reinforcement learning exploration is vital once formatting syntax is internalized. (iii) *The Decisive Role of Process Supervision*: Introducing verified process supervision (+SFT + Process-supervised RL, 50.15%) yields an additional +2.24% gain, boosting tasks demanding strict evidence verification such as CG-Bench (+4.31%) and SciVideoBench (+3.00%). On VideoTT (42.28% vs. 42.88%), outcome-only models occasionally exploit superficial single-frame cues in short clips, whereas our framework strictly enforces verified multi-interval alignment, ensuring robust long-horizon reasoning fidelity.

4.3 In-depth Analysis of Grounding and Captioning Capabilities

To validate that the performance gains of Video-Thinker stem from enhanced intrinsic "grounding" and "captioning" capabilities, we conduct a quantitative diagnostic comparison against the pre-trained base model Qwen3-VL-8B-Instruct and the representative post-trained video reasoning baseline VideoChat-R1.5 (Li et al., 2025b). We utilize 1K samples from ActivityNet (Caba Heilbron et al., 2015), YouCook2 (Zhou et al., 2018), and TutorialVQA (Colas et al., 2019), adopting the evaluation metrics detailed in Appendix F. Each sample contains one or multiple ground truth question-relevant key segment time annotations for grounding verification and corresponding ground truth captions for captioning evaluation.

Superior video grounding performance. To assess temporal localization, models output relevant intervals within `<time></time>` tags evaluated via mIoU (Yu et al., 2016) and Recall@K (Patel et al., 2022). As reported in Table 3, Video-Thinker substantially outperforms both baselines across all metrics, achieving 52.74% mIoU (+20.29% over base, +11.06% over VideoChat-R1.5) and 58.35% on strict Recall@0.5 (nearly doubling the base model's 29.14%). Overall, Video-Thinker reaches 65.24% average (+82.2% relative gain over base), confirming that process supervision over certified evidence intervals actively sharpens temporal boundary sensitivity.

Superior video captioning performance. To evaluate descriptive perception, models generate descriptions for video segments evaluated by BLEU@1 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-L (Lin, 2004). As shown in Table 3, Video-Thinker establishes the highest quality across all metrics (17.85% METEOR, 22.94% ROUGE-L, 17.91% BLEU@1, 19.57% average), consistently outperforming Qwen3-VL-8B-Instruct (+32.6% relative gain) and VideoChat-R1.5 (16.65%). These results demonstrate a mutual synergy: precise temporal grounding isolates informative visual frames for captioning, while descriptive captions ground the semantic rationale for deduction.

We further validate that Video-Thinker's endogenous reasoning paradigm outperforms modular external tool-use architectures—including plug-and-play tools, self-tool invocation, and agent frameworks—across all three tested benchmarks, while delivering an estimated $3\times$–$4\times$ speedup over multi-agent pipelines. Full comparisons and efficiency analysis are provided in Appendix J.

5 Conclusion

We introduce Video-Thinker, a framework that advances multimodal models toward authentic Thinking with Videos by unifying video reasoning graph-based data synthesis with evidence-aware reinforcement learning, enabling intrinsic temporal reasoning without external tools. Broader limitations and future directions are discussed in Appendix K.

AI USE STATEMENT

During the preparation of this manuscript, generative AI tools were used solely for grammar, spelling, and stylistic polish. All conceptual formulations, methodological designs, experimental evaluations, and scientific analyses were conceived, verified, and written by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our research utilizes publicly available, open-source video benchmarks and datasets, with no human subjects or private data collection involved. We anticipate that our work will contribute positively to transparent, trustworthy, and efficient video reasoning.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, the detailed algorithmic formulations, training hyperparameters, and evaluation protocols are documented in Section 3 and Section 4. Full implementation details, prompts, and benchmark specifications are provided in the Appendix, and our complete codebase is included in the supplementary material.

REFERENCES

James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26 (11):832–843, 1983.

Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint*, 2026.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024.

Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Bohan Zeng, Yang Shi, Sihan Yang, Pengfei Wan, Qiang Liu, Liang Wang, and Tieniu Tan. Versavid-r1: A versatile video understanding and reasoning model from question answering to captioning tasks. *arXiv preprint arXiv:2506.09079*, 2025b.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025c.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning, 2025d. URL https://arxiv.org/abs/2506.16141.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. Tutorialvqa: Question answering dataset for tutorial videos. *arXiv preprint arXiv:1912.01046*, 2019.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. *arXiv preprint arXiv:2505.24718*, 2025.

Google DeepMind. Gemini 3.1: Multimodal reasoning and flash generation. `https://blog.google/technology/google-deepmind/`, 2026.

Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *arXiv preprint arXiv:2510.08559*, 2025.

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.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

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.

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.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14953–14962, 2023.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Chenglin Li, Feng Han, Yikun Wang, Ruilin Li, Shuai Dong, Haowen Hou, Haitao Li, Qianglong Chen, Feng Tao, Jingqi Tong, et al. Videopro: Adaptive program reasoning for long video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7497–7513, 2026.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025a.

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*, 2025b.

Yunxin Li, Xinyu Chen, Zitao Li, Zhenyu Liu, Longyue Wang, Wenhan Luo, Baotian Hu, and Min Zhang. Veripo: Cultivating long reasoning in video-llms via verifier-gudied iterative policy optimization. *arXiv preprint arXiv:2505.19000*, 2025c.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Transactions on Multimedia*, 26:5548–5560, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

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*, 2025a.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning. *arXiv preprint arXiv:2505.14246*, 2025b.

David Ma, Yuanxing Zhang, Jincheng Ren, Jarvis Guo, Yifan Yao, Zhenlin Wei, Zhenzhu Yang, Zhongyuan Peng, Boyu Feng, Jun Ma, et al. Iv-bench: A benchmark for image-grounded video perception and reasoning in multimodal llms. *arXiv preprint arXiv:2504.15415*, 2025.

Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, et al. Taco: Learning multi-modal action models with synthetic chains-of-thought-and-action. *arXiv preprint arXiv:2412.05479*, 2024.

Muhammad Maaz, Hanoona Rasheed, Fahad Shahbaz Khan, and Salman Khan. Video-r2: Reinforcing consistent and grounded reasoning in multimodal language models. *arXiv preprint arXiv:2511.23478*, 2025.

OpenAI. Image thinking: Breakthroughs in visual chain-of-thought reasoning with OpenAI o3 and o4-mini. *OpenAI Blog*, April 2024. URL `https://openai.com/research/imagethinking`. Accessed: 2025-08-09.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Yash Patel, Giorgos Tolias, and Jiří Matas. Recall@ k surrogate loss with large batches and similarity mixup. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7502–7511, 2022.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *CoRR*, 2024a.

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

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. *arXiv preprint arXiv:2411.16044*, 2024.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, et al. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. *arXiv preprint arXiv:2505.10557*, 2025a.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in neural information processing systems*, 38:4350–4376, 2026.

Qunzhong Wang, Jie Liu, Jiajun Liang, Yilei Jiang, Yuanxing Zhang, Jinyuan Chen, Yaozhi Zheng, Xintao Wang, Pengfei Wan, Xiangyu Yue, et al. Vr-thinker: Boosting video reward models through thinking-with-image reasoning. *arXiv preprint arXiv:2510.10518*, 2025b.

Xiaodong Wang and Peixi Peng. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video, 2025.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025c.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*, 2025d.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28114–28128, 2025e.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478*, 2025.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

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*, 2025a.

Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520, 2016.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, et al. Vrbench: A benchmark for multi-step reasoning in long narrative videos. *arXiv preprint arXiv:2506.10857*, 2025b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20626–20636, 2025a.

Zeyu Zhang, Zijian Chen, Zicheng Zhang, Yuze Sun, Yuan Tian, Ziheng Jia, Chunyi Li, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Puzzlebench: A fully dynamic evaluation framework for large multimodal models on puzzle solving. *arXiv preprint arXiv:2504.10885*, 2025b.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10665–10674. IEEE, 2020.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1702–1713, 2025.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

A Extended Related Work

Thinking with Images. Recent breakthroughs in reinforcement learning-based post-training have demonstrated remarkable improvements in complex reasoning, exemplified by OpenAI-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025). Building upon this foundation, Multimodal Large Language Models (MLLMs) are undergoing a paradigmatic shift from treating visual inputs as passive, static context (Su et al., 2025) toward elevating visual operations into active cognitive tools (OpenAI, 2024). In this emerging paradigm, models dynamically inspect, manipulate, and interleave visual observations directly within their reasoning traces. For instance, Deepeyes (Zheng et al., 2025) enables models to autonomously invoke visual operators (such as zoom and cropping) during chain-of-thought reasoning via reinforcement learning. Similarly, Visual-ARFT (Liu et al., 2025b) utilizes GRPO (Shao et al., 2024b) to cultivate strategic planning and stepwise visual tool use through executable Python operators. These efforts demonstrate that transforming visual inspection into an explicit, manipulable reasoning step substantially enhances spatial problem-solving fidelity.

Reinforcement Learning for Video Reasoning. Extending these reasoning capabilities to dynamic videos represents the natural frontier for MLLMs, transitioning from static spatial perception toward capturing the temporal and logical structures of continuous events (Wang & Peng, 2025; Dang et al., 2025; Yu et al., 2025a). To cultivate video reasoning, recent efforts have explored reinforcement learning-based post-training. Video-R1 (Feng et al., 2025) applies GRPO to video domains to elicit implicit spatiotemporal reasoning, yet demands massive corpora exceeding 165K instances. VideoChat-R1 (Li et al., 2025b) incorporates reinforcement fine-tuning to balance spatiotemporal grounding with conversational capabilities. Temporal-R1 (Li et al., 2025a) introduces temporal grounding rewards alongside variance-aware data filtering to improve reasoning efficiency. Despite these advances, existing post-training methods predominantly rely on coarse outcome rewards over final answers (Shao et al., 2024b), which inadvertently reinforce superficial shortcuts and lucky guesses with hallucinated timestamps. Moreover, existing training data lacks structural guarantees that questions strictly necessitate temporal evidence, leaving models prone to text-only biases.

Intrinsic Reasoning vs. External Tools. To navigate long video sequences, several approaches deploy external auxiliary modules or pipelined tools (Liu et al., 2025a). However, such modular frameworks suffer from cascading errors across isolated components and contextual disconnection during iterative tool calls. In contrast, Video-Thinker advances an endogenous paradigm: extending "Thinking with Images" to "Thinking with Videos" by directly internalizing temporal localization (`<time>`) and visual description (`<caption>`) into the model's native chain of thought. By synthesizing certified tasks over video reasoning graphs and providing verified temporal process supervision during policy optimization, Video-Thinker establishes robust, faithful video reasoning without relying on external tool invocation.

B Overall Algorithm of Video-Thinker

Algorithm 2 provides the complete procedural pipeline of Video-Thinker, detailing the progression from video reasoning graph construction and task synthesis to two-stage policy training with process supervision.

C Source Datasets and Data Statistics

To construct our training corpora of 20K instances (10K for SFT cold-start and 10K for process-supervised RL), we source dense, high-quality event segment annotations across four prominent open-source video corpora, comprising a total of 22,948 unique untrimmed videos (Figure 2, Figure 3, and Table 4). These datasets collectively encompass three distinct video reasoning domains with varied temporal complexity profiles:

- **MDVC** (Krishna et al., 2017) (*Daily Activities*): Sourced from the dense multimodal video captioning benchmark on ActivityNet, containing 9,123 untrimmed videos (39.76% of the total corpus). It captures diverse open-domain human activities with temporally dense interval annotations, featuring an average of 3.73 event segments per video (median 3.0), providing concise and fine-grained temporal localization foundations.

Algorithm 2 VIDEO-THINKER Training and Synthesis Pipeline

Input: Source untrimmed videos $\mathcal{V}_{\text{raw}}$ with event descriptions, pre-trained base model π_θ (QWEN3-VL-8B-INSTRUCT), group size $G = 8$, trade-off $\lambda \leq 1/\sqrt{G-1}$.
Output: Optimized VIDEO-THINKER model policy π_θ.

 # Phase 1: Video Reasoning Graph Construction & Task Synthesis (§3.1, §3.2)
for each video $V \in \mathcal{V}_{\text{raw}}$ **do**
  Extract event segments $\{v_i = (I_i, c_i)\}_{i=1}^{n}$ with continuous intervals $I_i = [t_i^s, t_i^e]$ and descriptions c_i.
  Compute deterministic temporal edges $\mathcal{E}_{\text{temp}}$ across all pairs via Allen's 13 interval relations.
  Predict directed propositional entailment edges $\mathcal{E}_{\text{ent}}$ via single-step atomic fact verification.
  Collapse mutually entailing events ($u \Rightarrow v \wedge v \Rightarrow u$) into canonical event set $\hat{\mathcal{V}}$.
  Construct video reasoning graph $\mathcal{G} = (\hat{\mathcal{V}}, \mathcal{E}_{\text{temp}}, \mathcal{E}_{\text{ent}})$.
  Sample pairwise non-entailing evidence subset $A = \{s_1, \ldots, s_k\} \subset \hat{\mathcal{V}}$ satisfying $\forall u \neq v \in A$, $\neg(u \Rightarrow v) \wedge \neg(v \Rightarrow u)$.
  Formulate question Q, ground-truth deduction $y = \mathcal{R}_Q(A)$, and distractors $\{d_i\}_{i=1}^{k}$ where $d_i = \mathcal{R}_Q(A \setminus \{s_i\})$.
  Generate structured reasoning trajectory T with `<time>` and `<caption>` tags.
  Register certified evidence spans $\mathcal{E}^* = \{\mathcal{I}(s_j)\}_{j=1}^{k}$ where $\mathcal{I}(s_j) = \bigcup_{u \in \hat{\mathcal{V}},\, u \Rightarrow s_j} \mathcal{T}(u)$.
  Add (V, Q, T, y) to $\mathcal{D}_{\text{SFT}}$ (10K) and $(V, Q, y, \mathcal{E}^*)$ to $\mathcal{D}_{\text{RL}}$ (10K).
end for
 # Phase 2: SFT Cold-Start for Reasoning Syntax (§3.3)
for each batch in $\mathcal{D}_{\text{SFT}}$ **do**
  Compute cross-entropy loss $\mathcal{L}_{\text{SFT}}(\theta) = -\sum_t \log p_\theta([T; y]_t \mid V, Q, [T; y]_{<t})$.
  Update θ via gradient descent to instill `<time>` and `<caption>` formatting priors.
end for
 # Phase 3: Evidence-Aware Policy Optimization via GRPO (§3.3)
for each iteration with $(V, Q, y, \mathcal{E}^*) \in \mathcal{D}_{\text{RL}}$ **do**
  Sample G candidate reasoning trajectories $\{T^{(1)}, \ldots, T^{(G)}\} \sim \pi_{\theta_{\text{old}}}(T \mid V, Q)$.
  Parse predicted intervals $C^{(i)}$ from `<time>` tags of each candidate $T^{(i)}$.
  Compute multi-span evidence score $e^{(i)} = \min_j \max_{I \in \mathcal{I}(s_j)} \max_{c \in C^{(i)}} \text{tIoU}(c, I)$.
  Compute standardized outcome feedback $z(a^{(i)})$ and mean-centered evidence $(e^{(i)} - \bar{e})$.
  Formulate advantage $A_i = z(a^{(i)}) + \lambda(e^{(i)} - \bar{e})$ under boundary condition $\lambda \leq 1/\sqrt{G-1}$.
  Update θ by maximizing GRPO clipped objective $\mathcal{J}_{\text{GRPO}}(\theta)$.
end for
return Optimized policy π_θ.

- **VideoXum** (Lin et al., 2023) (*Event Summaries*): Sourced from the video summarization benchmark on ActivityNet, comprising 7,980 untrimmed videos (34.77%). It provides comprehensive narrative summaries with an average of 3.74 key event segments per video (median 3.0), offering dense event progression and transition supervision.

- **VidSTG** (Zhang et al., 2020) (*Spatio-Temporal Interaction*): Sourced from the VidOR video object relation benchmark, encompassing 4,055 untrimmed videos (17.67%). Characterized by complex multi-object spatial interactions and intricate temporal dynamics, it exhibits a heavy-tailed segment distribution with an average of 8.27 segments per video (median 6.0, spanning up to 70 segments), establishing challenging multi-entity, long-horizon temporal dependencies.

- **YouCook2** (Zhou et al., 2018) (*Cooking Procedures*): Comprises 1,790 instructional culinary videos from YouTube (7.80%). It features structured step-by-step procedure intervals with an average of 7.73 segments per video (median 7.0), furnishing coherent multi-stage causal and procedural temporal chains.

As visualized in Figure 3 and summarized in Table 4, the four corpora exhibit a distinct temporal complexity hierarchy: MDVC and VideoXum focus on concise, dense daily activities ($\mu \approx 3.7$ segments), YouCook2 features structured procedural steps ($\mu = 7.7$ segments), and VidSTG exhibits a heavy-tailed complexity profile with up to 70 segments per video ($\mu = 8.3$), providing the structural foundation for constructing multi-hop video reasoning graphs.

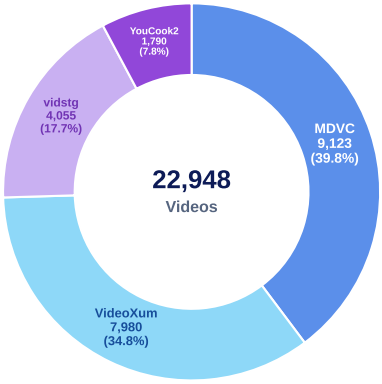

Figure 2: Sample breakdown and proportion across the four source video corpora.

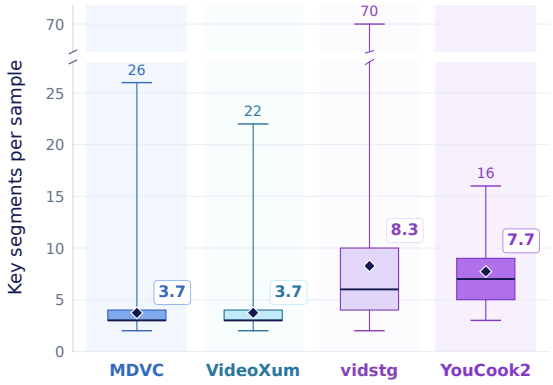

Figure 3: Distribution of annotated key temporal segments per video across the four source datasets.

Table 4: Detailed statistics and complexity distributions of the four source video datasets. All statistics are computed over filtered, verified event sequences used for video reasoning graph generation. Overall mean (4.85) and median (3.0) represent sample-weighted statistics across all 22,948 videos.

Source Dataset	Domain	Base Corpus	# Videos	Ratio	Mean (μ)	Median	Range
MDVC (Krishna et al., 2017)	Daily Activities	ActivityNet	9,123	39.76%	3.73	3.0	[2, 26]
VideoXum (Lin et al., 2023)	Event Summaries	ActivityNet	7,980	34.77%	3.74	3.0	[2, 22]
VidSTG (Zhang et al., 2020)	Video Interactions	VidOR	4,055	17.67%	8.27	6.0	[2, 70]
YouCook2 (Zhou et al., 2018)	Cooking Procedures	YouTube	1,790	7.80%	7.73	7.0	[3, 16]
Total / Overall	Diverse Real-World	Multi-source	**22,948**	**100.00%**	**4.85**	**3.0**	**[2, 70]**

By unifying these complementary sources—ranging from focused short-horizon daily activities (MDVC, VideoXum) to extended multi-agent interactions (VidSTG) and procedural workflows (YouCook2)—into structured video reasoning graphs, VIDEO-THINKER synthesizes certified reasoning tasks that demand genuine multi-segment temporal observation.

D BENCHMARK DATASETS

We evaluate VIDEO-THINKER across six challenging video reasoning benchmarks:

Video-Holmes (Cheng et al., 2025) probes narrative-driven deduction over 270 mystery and detective films accompanied by 1.8K questions. Resolving tasks requires multi-clue integration, causal discovery, and social relationship reasoning across complex video storylines, setting a stringent test for higher-order cognitive capabilities.

CG-Bench-Reasoning (Chen et al., 2024) comprises 1.2K long-form videos and 12K questions designed to evaluate clue-grounded video perception and reasoning. We evaluate specifically on the reasoning split, which demands explicit evidence retrieval and multi-step deduction, penalizing superficial text shortcuts.

VRBench (Yu et al., 2025b) evaluates multi-step temporal reasoning across 1,010 narrative videos in eight languages. It provides fine-grained annotations of reasoning steps and a multi-phase evaluation protocol, testing whether models can maintain faithful temporal grounding while solving complex queries.

SciVideoBench (Deng et al., 2025) evaluates multimodal scientific video reasoning through 1,000 challenging multiple-choice questions curated from over 240 scientific experiment videos across 25 disciplines. It tests whether models can integrate domain-specific knowledge with complex temporal dynamics.

VideoTT (Zhang et al., 2025a) benchmarks video interpretation correctness and robustness across 5,000 questions situated in 1,000 YouTube Shorts. It probes narrative complexity through adversarial questions, measuring model resilience against hallucination in fast-paced videos.

VideoMME (Fu et al., 2024) provides an expansive multi-modal evaluation suite featuring 2,700 questions over 900 videos spanning short, medium, and long durations (up to 1 hour). It comprehensively measures multimodal perception, temporal comprehension, and high-level reasoning.

E BASELINE MODELS

We compare VIDEO-THINKER against the pre-trained base model and 17 post-trained reasoning models reported in Table 1:

Pre-trained Base Model. Qwen3-VL-8B-Instruct (Bai et al., 2026) serves as our foundational multimodal backbone. It features native dynamic resolution vision encoding, advanced temporal patch modeling, and competitive multimodal instruction-following capabilities.

Post-trained Reasoning Models.

- **Temporal-R1-7B** (Li et al., 2025a): Post-trained via reinforcement learning with explicit temporal grounding rewards and variance-aware data filtering to improve spatiotemporal reasoning.
- **Open-R1-Video-7B** (Wang & Peng, 2025): Adapts the R1 reinforcement learning paradigm to video reasoning using GRPO with outcome-level correctness rewards.
- **TW-GRPO-7B** (Dang et al., 2025): Employs a token-weighted GRPO formulation with soft reward shaping for fine-grained multimodal reasoning.
- **Video-R1-7B** (Feng et al., 2025): Post-trains Qwen2.5-VL-7B using GRPO over a large-scale corpus of 165K multimodal instances to incentivize self-reflective reasoning.
- **Time-R1-7B** (Wang et al., 2025c): Incorporates temporal boundary rewards to enhance event prediction and retrospective analysis in long video contexts.
- **VideoChat-R1-7B** and **VideoChat-R1-Thinking-7B** (Li et al., 2025b): Integrate structured spatiotemporal grounding supervision with conversational alignment to bolster dynamic video dialogue.
- **GRPO-CARE-7B** (Chen et al., 2025d): Introduces coherence-aware reward modeling into GRPO to align intermediate reasoning steps with final answer choices.
- **VersaVid-R1-7B** (Chen et al., 2025b): Addresses conflicting objectives between captioning and question answering through intermediate proxy tasks.
- **VideoRFT-7B** (Wang et al., 2026): Employs reinforcement fine-tuning over multi-expert curated traces with semantic consistency rewards across 310K training samples.
- **VR-Thinker-7B** (Wang et al., 2025b): Adapts thinking-with-images operations to video understanding via rejection sampling and outcome-based GRPO over 20K samples.
- **Video-RTS-7B** (Wang et al., 2025e): Utilizes pure reinforcement learning with video-adaptive test-time scaling to optimize temporal sampling efficiency.
- **Video-R2** (Maaz et al., 2025): An advanced video reasoning model optimized through reinforcement learning with multi-turn temporal exploration.
- **VideoChat-R1.5** (Yan et al., 2026): Enhances VideoChat-R1 with scaled multimodal data (80K instances) and refined spatiotemporal reward mechanisms.
- **Video-MTR** (Xie et al., 2025): Optimizes multi-task temporal reasoning through task-adaptive reward modeling over an 8K curated training corpus.
- **VerlPO-7B** (Li et al., 2025c): Implements proximal policy optimization over 117K video reasoning instances using verifier-guided iterative reinforcement learning.
- **VideoPro-GRPO** (Li et al., 2026): Employs adaptive program reasoning and process reward shaping within an 8B vision-language backbone to supervise multi-segment video interpretation.

F Evaluation Metrics

Accuracy. For all question-answering benchmarks (Video-Holmes, CG-Bench-Reasoning, VRBench, SciVideoBench, VideoTT, VideoMME), we measure standard top-1 accuracy (%). Predictions are parsed from the `<answer>` tags and matched deterministically against ground-truth choices.

Temporal Grounding Metrics. To evaluate temporal localization fidelity, we adopt:

- **Mean Intersection-over-Union (mIoU)** (Yu et al., 2016): Measures the temporal overlap between predicted intervals $p = [t_s^p, t_e^p]$ and ground-truth intervals $g = [t_s^g, t_e^g]$, computed as $\text{IoU} = \frac{|p \cap g|}{|p \cup g|}$ and averaged across all test instances.
- **Recall@K** (Patel et al., 2022): Measures the proportion of ground-truth segments successfully retrieved with IoU $\geq K$. We report Recall@0.3 (coarse coverage) and Recall@0.5 (strict boundary alignment).

Video Captioning Metrics. To assess the descriptive accuracy of visual clues, we employ:

- **BLEU@1** (Papineni et al., 2002): Quantifies unigram lexical precision between predicted captions and ground-truth descriptions.
- **METEOR** (Banerjee & Lavie, 2005): Computes the harmonic mean of precision and recall with explicit word alignment, synonym matching, and stemming penalties.
- **ROUGE-L** (Lin, 2004): Evaluates sentence-level structure correspondence based on the Longest Common Subsequence (LCS).

G Prompts

G.1 Evaluation Prompt Template

During inference across all benchmarks, models are queried with the unified prompt template:

Prompt Template for Evaluation

System Prompt: You are an expert video analyst tasked with solving problems based on video content. When answering a question about a video, you should carefully observe and analyze important visual clues from the video. Throughout your reasoning process, you should actively identify key video segments: specify each relevant time range with `<time>start_time-end_time</time>`, and describe the observed visual clues with `<caption>Description of visual clues</caption>`. After examining the necessary visual clues, synthesize your observations to determine the final answer. Finally, conclude by placing your final answer in `<answer> </answer>` tags.
Question Template: {Question}
Please analyze the video carefully, identifying key video segments and their visual clues using `<time> </time>` and `<caption> </caption>` tags. Then conduct analysis and reasoning to arrive at your answer to the question. Finally, provide only the single option letter (e.g., A, B, C, D) within the `<answer> </answer>` tags.

G.2 Propositional Entailment Labelling Prompt

As detailed in Section 3.1, for each untrimmed video, we construct directed propositional entailment edges $\mathcal{E}_{\text{ent}}$ over candidate event pairs using Gemini-3.1-Flash (DeepMind, 2026). The entire set of candidate pairs for a video is evaluated in a single API call to prevent incomplete relation graphs. The system prompt instructs the model to evaluate entailment via an atomic factual checklist rather than subjective impressions, restricting inference to at most a single unambiguous step. Non-entailing relations (such as causal likelihood, common precedence, or different occurrences of identical actions) are strictly excluded. The complete prompt template is presented below:

Propositional Entailment Labelling Prompt

System Prompt: You compare short descriptions ("captions") of moments in one video and decide which ones entail which.

Write `A => B` when observing only A already lets you conclude that whatever B describes has happened. A is the more specific description; B says strictly less, and everything B says is already said by A.
Decide it as a checklist, not an impression. Break B into the atomic facts it asserts—who, what action, on what object, with what attribute—and check each one off against A:

- A: *"a man puts a plate into the sink"*, B: *"a man has held a plate"*
  → **who**: a man (A says it); **what**: has held it (putting it into the sink requires having held it → A supports it); **object**: a plate (A says it); B asserts nothing else $\implies$ $\mathbf{A} \Rightarrow \mathbf{B}$.
- A: *"a man holds a cup filled with water"*, B: *"a man holds a cup"*
  → **who**: a man (A says it); **action**: holds (A says it); **object**: a cup (A says it); B asserts nothing else $\implies$ $\mathbf{A} \Rightarrow \mathbf{B}$.

If even one fact in B is missing from A, or contradicted by A, the answer is no. Only one step of inference is allowed—the kind a person would make instantly and without hesitation (holding a full cup means holding a cup; putting a plate down means having held it). If establishing B from A takes a chain of such steps, or requires assuming what usually happens rather than what must be true, treat it as not entailed.

These are NOT entailment, and this is the more common mistake to guard against—do not report a relation just because one thing plausibly leads to or is likely alongside the other:

- *"a man holds a knife"* vs. *"a man is cutting vegetables"*: Holding a knife does not mean cutting is happening; it is merely possible. This is a causal / likelihood relation, not entailment. **No edge.**
- *"a man walks into the kitchen"* vs. *"a man opens the fridge"*: One often precedes the other, but neither statement guarantees the other occurred. **No edge.**
- *"two people are talking"* vs. *"a woman is talking"*: Neither refines the other: the first does not say who, the second does not say how many are present. **No edge.**
- *"he hits the bag" at 0:01* vs. *"he hits the bag" at 0:30*: Identical wording, but two separate occurrences. An event does not entail a different occurrence of itself, however similarly worded. **No edge.**

If both directions hold—the two captions state the same fact in different words—report both; they will be merged into one node downstream, which is what makes the relation a strict partial order rather than a preorder.
Judge only from the captions given. Do not use what the video probably showed beyond what a caption states. Most pairs in a video are unrelated or merely likely, not entailing—returning few or no edges is the normal, expected outcome, not a sign you are being too strict.
Return strictly valid JSON and nothing else:

```
{"edges": [{"specific": 3, "general": 1}]}
```

`"specific"` is the event whose caption entails the other; `"general"` is the one it entails. Omit every pair you reject; most pairs should be omitted.

User Prompt Template:

```
[Events in this video]
- Event 1 [0:05-0:12]: "a man walks into the kitchen"
- Event 2 [0:12-0:18]: "a man holds a cup filled with water"
...
[Pairs to judge]
- Event 1 vs Event 2
...
```

For each pair, report an edge only if one event's caption entails the other by the definition above. Omit the rest.

H QUALITATIVE CASE STUDIES

To provide concrete insight into how VIDEO-THINKER performs intrinsic temporal reasoning across diverse video scenarios, we present five qualitative case studies from the evaluation benchmarks comparing VIDEO-THINKER against the base model Qwen3-VL-8B-Instruct:

- **Long-Horizon Event Localization and Narrative QA (Figure 4):** In long-form video question answering requiring navigation across an 8-minute clip, VIDEO-THINKER accurately pinpoints the critical rescue interval (`07:39-08:09`) to determine the woman's outcome, whereas the base model hallucinates an earlier drowning event due to lack of temporal grounding.
- **Multi-Occurrence Temporal Action Disambiguation (Figure 5):** In multi-occurrence actions, VIDEO-THINKER sequentially localizes the first and target second lighting events via `<time>` and

<caption> tags, whereas the baseline hallucinates timestamps during an inactive state, yielding a wrong answer.

- **Multi-Event Sequential Ordering and Entity Identification (Figure 6):** Probing complex temporal order (identifying the second animal to cut bridge ropes), VIDEO-THINKER sequentially grounds the first event (heron at 05–06s) and second event (monkey at 09–10s) within its reasoning trace, deducing the correct answer. In contrast, the base model fails to ground intermediate occurrences and outputs a flawed deductive guess.
- **Contiguous Temporal Action Disambiguation (Figure 7):** In videos featuring temporally contiguous human behaviors (swearing at 02:52–03:09 immediately followed by leaving at 03:10–03:30), VIDEO-THINKER accurately discriminates the boundary between contiguous actions and justifies its response through grounded visual descriptions.
- **Cross-Interval Multi-Clue Relational Deduction (Figure 8):** Deducing subtle real-world relationships between characters requires integrating distant clues across disparate intervals (reading a message at 00:30–00:44 and a face-to-face conversation at 04:27–04:37). VIDEO-THINKER successfully identifies both necessary evidence segments to establish their true relationship, avoiding superficial text-prior shortcuts.

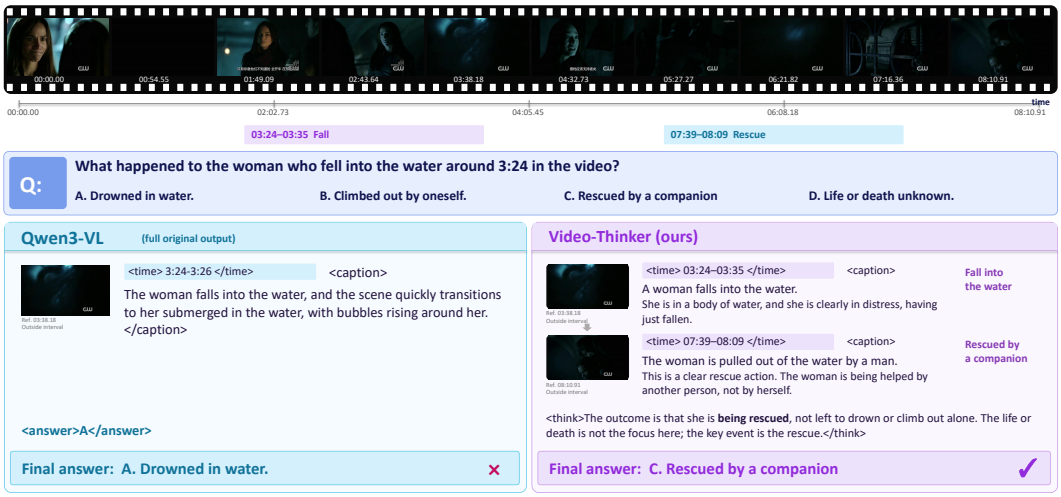

Figure 4: Qualitative comparison on long-horizon narrative reasoning. VIDEO-THINKER accurately localizes distant temporal evidence to resolve the query, while the baseline model suffers from temporal hallucination.

I BOUNDARY CONDITION ON EVIDENCE WEIGHT λ

To ensure that process supervision does not penalize factually correct answers, we examine the boundary condition on parameter λ. For a group of G sampled candidates where k responses are correct ($a = 1$) and $G - k$ are incorrect ($a = 0$), the sample mean is $\mu_a = k/G$ and standard deviation is $\sigma_a = \sqrt{\frac{k(G-k)}{G^2}}$. Standardizing the outcome yields an advantage for correct responses of $z_{\text{correct}} = \frac{1-k/G}{\sigma_a} = \sqrt{\frac{G-k}{k}}$, which attains its infimum at $k = G - 1$ (when all but one candidate are correct) with $z_{\min} = \frac{1}{\sqrt{G-1}}$. Since the mean-centered evidence deviation satisfies $|e^{(i)} - \bar{e}| < 1$, setting:

$$\lambda \leq \frac{1}{\sqrt{G-1}} \tag{9}$$

guarantees that $A_i > 0$ strictly holds for all correct trajectories, keeping process feedback aligned with answer correctness without flipping the gradient direction of correct answers.

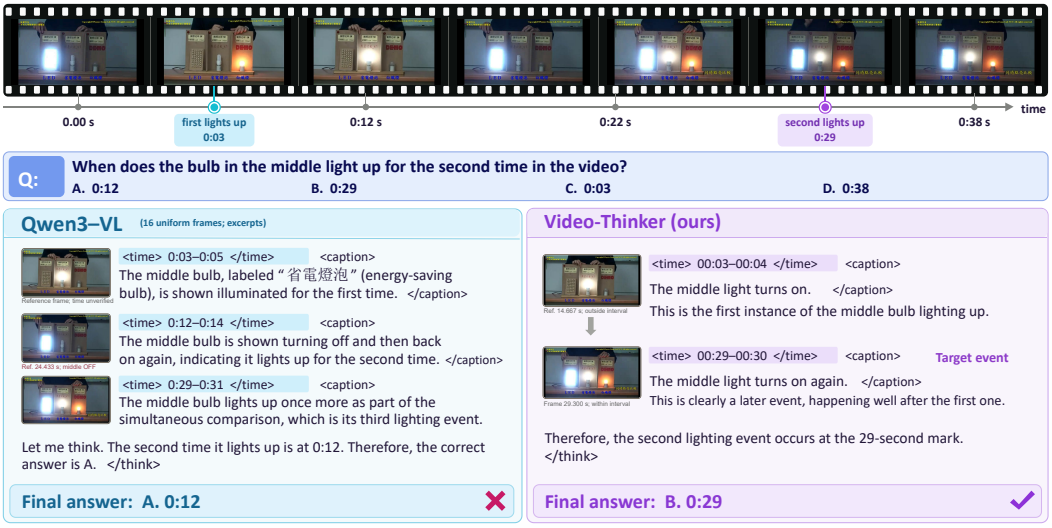

Figure 5: Qualitative comparison on multi-occurrence temporal action disambiguation. VIDEO-THINKER sequentially localizes the first and target second lighting events via <time> and <caption> tags, whereas the baseline hallucinates timestamps during an inactive state, yielding an erroneous conclusion.

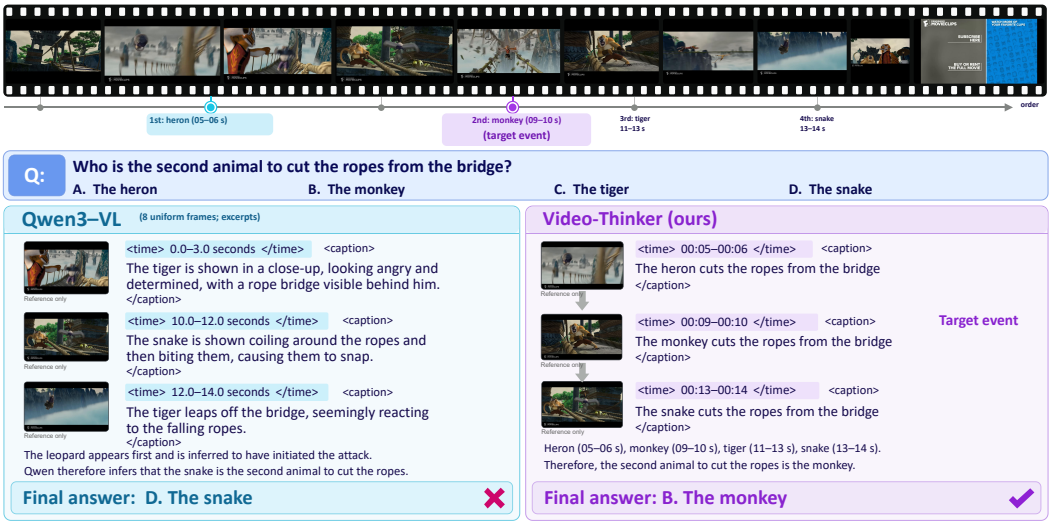

Figure 6: Qualitative comparison on multi-event temporal ordering. VIDEO-THINKER sequentially grounds each constituent event to determine chronological precedence, whereas the baseline fails to identify the intermediate action.

J ANALYSIS OF ENDOGENOUS REASONING VS. EXTERNAL TOOL CALLING

A central premise of this work is that cultivating intrinsic temporal grounding and captioning directly within native chain-of-thought processes outperforms modular external tool-use architectures. To validate our endogenous hypothesis, we compare VIDEO-THINKER against three tool-augmented paradigms on Video-Holmes, CG-Bench-Reasoning, and VRBench: (i) *Plug-and-Play Tools* (Temporal-R1-7B (Li et al., 2025a) for grounding and SkyCaptioner-V1-8B (Chen et al., 2025a) for captioning); (ii) *Self-Tool Invocation* (Qwen3-VL-8B-Instruct invoking itself in serialized multi-turn prompt calls); and (iii) *Agent Framework* VideoMind-7B (Liu et al., 2025a).

Performance Comparison. As reported in Table 5, incorporating external plug-and-play tools degrades the reasoning accuracy of the base model across all benchmarks (Video-Holmes drops from 40.39% to 36.84%, VRBench from 59.40% to 53.17%), stemming from representation disparity and



Figure 7: Qualitative comparison on contiguous temporal action disambiguation. VIDEO-THINKER precisely delineates adjacent temporal boundaries between successive human actions.

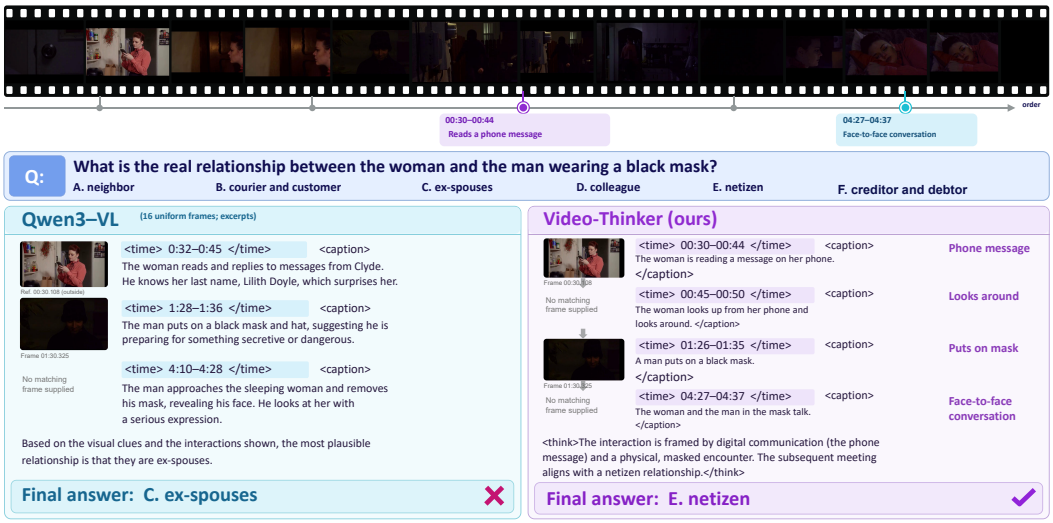

Figure 8: Qualitative comparison on cross-interval clue integration. Resolving the character relationship necessitates synthesizing visual evidence spanning distant temporal intervals, which VIDEO-THINKER achieves faithfully.

Table 5: Performance comparison of our model with methods with external tool use. G denotes "grounding tool" and C denotes "captioning tool". The best results are marked in **red bold**.

Model	Video-Holmes	CG-Bench	VRBench
Base Model			
Qwen3-VL-8B-Instruct	40.39%	36.66%	59.40%
Base Model + Plug-and-Play Tools			
G: Temporal-R1-7B C: SkyCaptioner-V1-8B	36.84%	32.48%	53.17%
G: Qwen3-VL-8B-Instruct C: Qwen3-VL-8B-Instruct	38.27%	34.18%	56.83%
Existing Tool-Use Framework			
VideoMind-7B	38.98%	31.99%	75.39%
Our Framework (Endogenous Reasoning)			
Video-Thinker (Ours)	**47.03%**	**41.56%**	**81.24%**

prompt distribution shift between specialized 7B modules and the 8B foundation. Self-tool invocation partially mitigates cross-model incompatibility (38.27% on Holmes, 34.18% on CG-Bench), but

still falls behind unassisted base reasoning (40.39%, 36.66%) due to pipeline serialization tax: early localization inaccuracies become rigid context priors that mislead downstream deduction. While VideoMind-7B achieves 75.39% on procedural tracking (VRBench), its rigid planning incurs sharp drops on complex narrative deduction (38.98% on Holmes). In contrast, VIDEO-THINKER achieves decisive superiority across all benchmarks (47.03%, 41.56%, 81.24%), establishing that endogenous temporal reasoning eliminates modular interface friction.

Efficiency Analysis. Beyond accuracy, computational efficiency is a decisive advantage of our endogenous paradigm. Orchestrating external tool pipelines introduces severe latency penalties from repeated context re-encoding and inter-process communication. In contrast, VIDEO-THINKER executes end-to-end within a single unified generative pass, completing evaluation in just 1.6h on Video-Holmes, 2.6h on CG-Bench, and 4.8h on VRBench—delivering an estimated $3\times$ to $4\times$ speedup over modular multi-agent pipelines.

K LIMITATIONS

While VIDEO-THINKER substantially improves temporal reasoning fidelity and data efficiency, several limitations remain to be addressed in future investigations: First, our framework currently relies on offline event segment proposals to construct video reasoning graphs during data synthesis. While inference operates end-to-end without external tool assistance, scaling the synthesis pipeline to extremely long, unannotated video archives requires developing more automated, streaming graph parsing strategies. Second, the current formulation focuses primarily on visual and textual modalities. Extending video reasoning graphs to integrate auditory streams, speech transcripts, and ambient audio cues represents an important direction for holistic multimodal reasoning. Finally, our process supervision evaluates temporal interval overlap via temporal IoU against certified evidence spans. While effective at penalizing ungrounded hallucinations, developing denser semantic verifiers that jointly assess the factual alignment of intermediate descriptions during reinforcement learning warrants further study.