Incentivizing Vision Language Models to Search for Long Video Question Answering

Harsh Goel*, S P Sharan*, Sahil Shah, Minkyu Choi, Joungbin An, Kristen Grauman, and Sandeep P. Chinchali

The University of Texas at Austin, Austin, TX, USA

Abstract. We introduce `VSeek`, an agentic framework that transforms long-video question answering (LVQA) from a passive, single-pass perception task into a multi-turn retrieval process. `VSeek` utilizes a natural language-driven search to identify relevant context within long videos and is post-trained with reinforcement learning (RL) to jointly formulate targeted search queries and reason over retrieved clips for LVQA. While RL post-training has revolutionized reasoning in symbolic domains such as mathematics and code, its application to long-video understanding remains hindered by a lack of verified rewards. To ensure that the retrieved context is relevant, we propose a novel neuro-symbolic approach that bridges open-ended natural language with discrete visual verification. Specifically, complex user queries are compiled into formal temporal logic specifications for systematically decomposing natural language questions into a definitive checklist of required atomic visual primitives, such as key objects and activities, along with their temporal ordering. These systematically derived grounding events provide the critical feedback signal for RL post-training, enabling dense, verifiable rewards based on the successful retrieval of these specific visual elements rather than relying entirely on outcome-only answer accuracy. By explicitly optimizing for this verifiable evidence-seeking behavior, `VSeek` improves Pass@1 scores by up to 8% and Pass@4 scores by 15% on long-video understanding benchmarks compared to base models. We open-source our code at `https://utaustin-swarmlab.github.io/VSeek`.

Keywords: Video Question and Answering · Multimodal Reasoning · Vision Language Models · Reinforcement Learning

1 Introduction

When humans are tasked with answering questions pertaining to a two-hour documentary, they do not rely on a sparse, random sampling of frames; instead, they scrub through the timeline to identify relevant segments. Despite this, current post-training paradigms for video-language models [6,40,19,5] fundamentally train models over uniformly sampled frames from long videos. This is insufficient for long-video understanding since uniform sampling inevitably

* Equal Contribution.

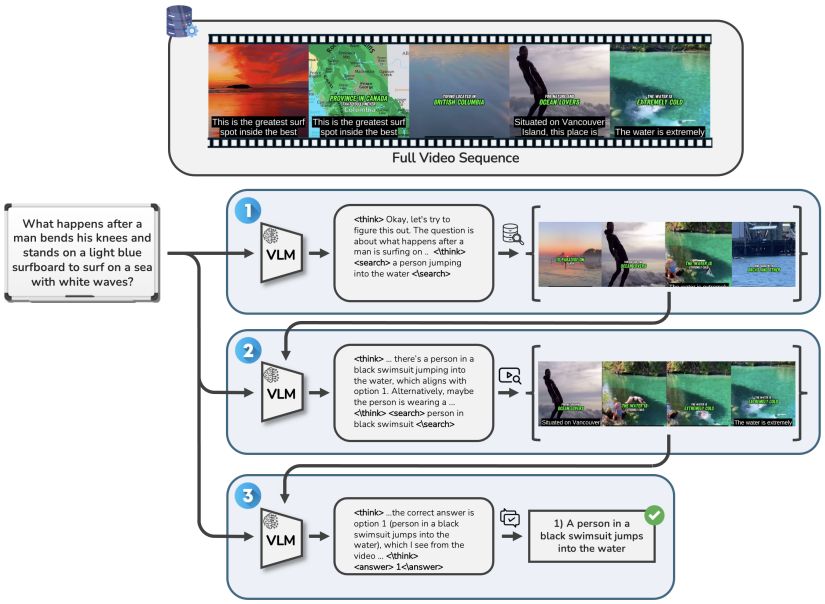

Fig. 1. Visualization of VSeek. Given an indexed long video and a question, `VSeek` iteratively decomposes the query, issues natural-language search requests to retrieve candidate clips, and refines its video context before producing the final answer.

misses the sparse, query-critical moments [55,69,74] required to answer complex queries while utilizing irrelevant context that leads to incorrect answers [27].

To tackle this challenge, recent work has explored acquiring relevant video context by equipping models with tools to retrieve specific clips [69,67,31,61,29,55], which entirely rely on prompting off-the-shelf foundation models to identify relevant frames from per-frame video captions. Other works have explicitly trained dedicated tool-calling vision-language models [26,77] to teach the model to identify and output precise time intervals to obtain relevant video clips through extensive supervised fine-tuning (SFT) before RL post-training. To bypass the limitations of exhaustive caption scanning and rigid temporal prediction, we propose `VSeek`, an agentic framework that actively searches for query-specific context with natural language as a primary search interface (see Figure 1). However, current pipelines rely almost exclusively on "outcome-only" supervision [6,5] alongside auxiliary objectives like temporal ordering [19,71], which would fail to certify whether an agent has constructed a query-specific context when using natural language-based video search.

We argue that the path forward for agentic long-video question answering (QA) must mirror the recent breakthroughs in large language model (LLM) reasoning for mathematics [56,57] and programming [50,23,63], where progress is catalyzed by verifiable rewards (*i.e.*, unit tests and formal symbolic rules). However, applying this level of exact rule-checking to raw video is difficult be-

cause natural language queries inherently entangle the semantic content of visual events with complex temporal dependencies. This is necessary to score the query-specific contexts built by video-understanding agents. Thus, in this work, we also introduce Visual Evidence via Temporal Logic (`VETL`), a pipeline to decouple visual evidence from its temporal dependencies in natural language questions, to formulate a reward signal. By translating these open-ended queries into formal temporal logic specifications, we systematically disentangle the underlying visual grounding events or primitives and represent their temporal relationships through temporal operators. Thus, long-horizon questions are transformed into an exact checklist of visual prerequisites (*i.e.*, entities, actions, and relations) that act as visual "unit tests" defining valid query-specific contexts. Through `VETL`, we obtain rewards for RL post-training for open-ended visual context construction, therefore improving the agentic search and answering process of `VSeek`.

Our contributions are summarized as follows:

- We introduce `VSeek`, an agentic framework that turns long-video QA into an active, multi-turn discovery process by decomposing complex queries and issuing targeted natural-language searches for relevant clips.
- We propose `VETL`, a neuro-symbolic pipeline that closes the verification gap by converting questions into video-grounded primitives and compiling them into verifiable temporal-logic specifications that formalize the "correct context" for any query.
- We design RL rewards from `VETL` that score an agent's evidence-gathering trajectory against the resulting extracted primitives, enabling checkable step-wise feedback beyond outcome-only answer supervision.
- We show that optimizing for verified context construction boosts long-horizon agentic reasoning: `VSeek` improves Pass@1 by up to 8% and Pass@4 by 15% over base models on long-video understanding benchmarks.

2 Related Work

2.1 Video Question Answering

Early visual question answering (VQA) systems [4,75] encoded entire videos through convolutional, recurrent or attention-based pipelines [30,81], but scaled poorly to longer inputs [28,62,9]. Modern vision-language models (VLMs) [1,6,38,11] enable zero-shot VQA with stronger generalization, yet these still operate over a fixed set of uniformly sampled frames. This dilutes relevant context in multi-event sequences [66,70,59,28] and degrades performance on queries requiring precise temporal localization.

Retrieval-augmented methods [29,67,69,74,25] mitigate this by conditioning on task-relevant segments, but typically rely on coarse heuristics or textual summaries that sacrifice visual fidelity. Dedicated tool-calling VLMs [26,77] offer more structured video grounding, yet their rigid interfaces demand extensive SFT to learn format-specific commands and timestamp mappings. `VSeek`

sidesteps both limitations by using flexible natural-language queries as the search interface, enabling an intuitive search without format-specific supervision.

2.2 Agentic Search and RL Post-Training

Text-based agentic frameworks [52,73,21,34] interleave reasoning with tool calls; systems such as Search-R1 [34] encapsulate thought chains and integrate retrieved evidence for better retrieval-augmented generation (RAG) [37]. These pipelines excel in knowledge-intensive text tasks, but extending them to video, where evidence is visual and not textual, remains unsolved. `VSeek` addresses this by treating long-video understanding as a temporal search-and-retrieval task.

RL post-training through Group Relative Policy Optimization (GRPO) [57,22] has driven rapid gains in math and code, where verifiable rewards (unit tests, compilation checks) provide dense and reliable training signals [15,23,63]. Transferring this to video is hard: existing pipelines use outcome-only supervision [77,19] or self-verification [47], and therefore do not directly incentivize correct context construction to prevent spurious shortcuts [3]. `VSeek` fills this gap by compiling queries into temporal-logic specifications [46,48] that act as visual "unit tests," providing rewards that incentivize correct evidence gathering rather than final-answer accuracy alone.

2.3 Neuro-Symbolic Methods for Video Understanding

Extracting symbolic representations from video objects, actions, relations, and events has been explored for classification [17,65], event detection [64,43], VQA [12], robotic planning [24,35], and autonomous driving verification [32,44]. These symbols are typically obtained from spatio-temporal graphs [51,76] or latent-space [8,36], and translated into formal languages such as temporal logic [7]. Within video understanding, symbolic methods have supported long-form reasoning [28,62], and recent work [72,13] represents videos as formal models to search for events. To represent queries for evaluation, grounding, and question-answering, recent work [55,45,58,14,41] extracts video-based primitives in temporal logic for subsequent neuro-symbolic verification. Crucially, all prior neuro-symbolic methods use temporal logic for *post-hoc* context construction; none use them as a *training signal* [54]. `VSeek` departs from this by compiling queries into formal specifications over visual primitives and using them to *shape dense rewards* during RL post-training, turning open-ended context construction into a verifiable objective.

3 Preliminaries

3.1 Search with LLMs

We draw inspiration from text-based search agents such as Search-R1 [34] and R1-Searcher [60], which auto-regressively generates a complete response $y \sim$

$\pi_\theta(\cdot \mid x)$, given a prompt x and a search engine, as an interleaved sequence of reasoning and tool calls. The policy π_θ decodes y autoregressively, so that the sequence-level distribution factorizes into per-token conditionals, $\pi_\theta(y \mid x) = \prod_t \pi_\theta(y_t \mid x, y_{<t})$; we use π_θ to denote both forms interchangeably throughout. Guided by a strict token protocol, these models reason within `<think>` tags, issue queries via `<search>`, and iteratively assimilate returned evidence or context within `<information>` tags until they produce a final answer within `<answer>` tags. We extend this paradigm to the video domain, transforming long-video understanding from a passive, single-pass perception problem into an **agentic retrieval** task. By explicitly interleaving active search with internal thinking, the model can dynamically construct context to produce cogent answers on longer videos.

3.2 Group Relative Policy Optimization for Multi-hop Search

RL methods optimize LLM policies with task-specific rewards. While Proximal Policy Optimization (PPO) [53] uses a learned value function, GRPO [57] eliminates its need by standardizing rewards within a sampled group.

The core principle of GRPO is to establish a relative baseline for policy updates using the average reward from a group of sampled trajectories. For each input prompt x from the dataset $\mathcal{D}$, GRPO samples a group of G responses $\{y_1, y_2, \ldots, y_G\}$ from a policy π_{old} to optimize the policy π_θ through the objective

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{\sum_{t:I(y_{i,t})=1}\Big(\mathcal{L}_{\text{clip}}(\theta,t)-\beta\mathcal{D}_{\text{KL}}(\theta,t)\Big)}{\sum_t I(y_{i,t})}\right], \tag{1}$$

where the clipped surrogate per-token loss $\mathcal{L}_{\text{clip}}(\theta, t)$ at position t is defined as

$$\mathcal{L}_{\text{clip}}(\theta,t) = \min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}\Big), \tag{2}$$

and the KL regularization $\mathcal{D}_{\text{KL}}(\theta, t)$ against the reference policy π_{ref} is

$$\mathcal{D}_{\text{KL}}(\theta,t) = \text{KL}\Big[\pi_\theta(\cdot \mid x, y_{i,<t})\Big\|\pi_{\text{ref}}(\cdot \mid x, y_{i,<t})\Big]. \tag{3}$$

Here, $r_{i,t}(\theta)$ denotes the probability ratio between the current and old policies, and $y_{i,<t}$ denotes tokens sampled up to the decoding step t. Hyperparameters ϵ and β control the clipping range and the KL-regularization strength, respectively. The advantage $\hat{A}_{i,t}$ is calculated based on the rewards within the group. For agentic search, a token mask $I(y_{i,t})$ is applied to ensure that the model does not learn to reproduce or memorize the outputs obtained from external retrieval tools.

3.3 Temporal Logic

Temporal logic (TL) is a formal language for describing how events evolve over time using logical and temporal operators. We illustrate temporal logic given

the following query and answer: *"After the woman pours hot water and spoons yogurt into the bowl, what does she place as a topping? Answer: Strawberry"*. Events are represented as atomic primitives that evaluate to `True` or `False` at each time step. These primitives are combined with the logical operators `AND` ($\wedge$), `OR` ($\vee$), `NOT` ($\neg$) and `IMPLY` ($\rightarrow$), and the temporal operators `ALWAYS` ($\Box$), `EVENTUALLY` ($\Diamond$), `NEXT` ($\bigcirc$) and `UNTIL` ($\mathcal{U}$).

Returning to our running example, we define a set of atomic primitives $\mathcal{P}$ and a corresponding TL specification φ as follows:

$$\mathcal{P} = \{\texttt{woman}, \texttt{pours water}, \texttt{spoons yogurt}, \texttt{adds strawberry}\},$$
$$\varphi = \Box\,(\texttt{woman}) \wedge ((\texttt{pours water} \wedge \texttt{spoons yogurt}) \wedge \Diamond\,(\texttt{adds strawberry}))\,.$$

This specification encodes the requirements of the context that `VSeek` would need to build. This allows us to verify whether the video context retrieves the correct evidence, providing a checkable objective for reinforcement learning.

4 Method

In this section, we introduce `VSeek`, an agentic framework designed to perform precise evidence seeking within long-form videos. We first formalize this interaction as a sequential decision process (§4.1). We then detail our tool infrastructure, which leverages both semantic visual embeddings and subtitle-based indexing to bridge the gap between natural language queries and temporal video segments (§4.2). Finally, to overcome the challenge of sparse rewards in reinforcement learning for video tasks, we present `VETL`, a neuro-symbolic reward shaping mechanism that provides dense, step-wise feedback based on the retrieval of verifiable visual primitives (§4.3).

4.1 Interactive Video-Evidence Seeking Trajectory

Following agentic search frameworks in LLMs, we formulate `VSeek` as a sequential decision-making process where a single VLM interacts with an environment via retriever-based tools.

Given an initial user prompt x and a long video V, the initial context of the agent is defined as $c_0 = x$. At each turn j, we autoregressively decode reasoning tokens a_j^{think} (within `<think>` and `</think>`), followed by a search action a_j formatted as `<search><args>...</args></search>` from the VLM policy π_θ. Therefore, the generated response y_j at turn j is $(a_j^{\text{think}}, a_j) \sim \pi_\theta(\cdot \mid c_{j-1})$ conditioned on an aggregated context c_{j-1}. Executing the search action a_j with the search engine yields a specific segment of frames $o_j \subset V$ retrieved from the source video V. The agent's context is then sequentially augmented with the generated tokens and the retrieved video clip as $c_j = [c_{j-1}, a_j^{\text{think}}, a_j, o_j]$. The agent's trajectory terminates at turn J when the VLM emits its final response within `<answer>` and `</answer>`. This yields a complete trajectory $\tau = (c_0, a_1^{\text{think}}, a_1, o_1, \ldots, a_J^{\text{think}}, a_J)$. Upon termination, the environment evaluates the response and assigns a terminal exact match reward $r_J \in \{0, 1\}$ (also

referred to as R_{em}) based on the final accuracy of the VQA. During post-training, the objective is to optimize π_θ to maximize the expected reward.

4.2 Agentic Design and Tool Infrastructure

We built `VSeek` around a VLM configured to seamlessly alternate between reasoning and evidence acquisition. A strict design constraint is that all tool outputs must fit within the VLM's visual context window; therefore, we enforce a per-turn budget of exactly 16 frames for observation o_t.

Preprocessing and Indexing. Before agent interaction, we preprocess the prescribed long video V by segmenting them into 8-second fixed-length clips. Each clip is embedded using the ViCLIP [68] video encoder referenced as e_{video}, whereas language queries and subtitles are embedded with ViCLIP's text encoder e_{text}. Although alternative temporal grounding models such as CLIP [49], LanguageBind [80], GroundingDino [42], or other event detection models [39,18,33] can, in principle, be employed, we adopt ViCLIP in this work owing to its demonstrated performance and inference efficiency. Concurrently, we index the video's subtitles, explicitly mapping each text transcript to the exact frame windows of its occurrence. This dual-index forms the foundation for the agent's retrieval mechanisms.

Actionable Tools. Following indexing, we define a suite of tools that allow the VLM to query the video index solely via natural language. The model invokes these tools using explicit XML-style structural tags:

1. **Search-based Tool:** Invoked via the `<search> search_string </search>` tag, this tool performs semantic visual retrieval by embedding the query with the ViCLIP text encoder and scoring it against pre-computed video clip embeddings. It returns the top-k clips $\mathcal{K}$ that maximize cosine similarity:
$$\mathcal{K} = \underset{v_i \in V}{\text{Top-}k} \left(\frac{e_{\text{text}}(\tilde{q}) \cdot e_{\text{video}}(v_i)}{\|e_{\text{text}}(\tilde{q})\| \|e_{\text{video}}(v_i)\|} \right), \tag{4}$$
where $\tilde{q}$ is the `search_string`, and v_i is the i-th clip in the index. To satisfy the constraint of the visual context, the environment uniformly samples frames from clips in $\mathcal{K}$ to form the observation o_t, capped at 16 frames.
2. **Subtitle-based Search:** This tool searches the video index by subtitle using the `<search_subtitle> search_string </search_subtitle>` tag. It bypasses the visual embedding space and directly matches the text query $\tilde{q}$ to the indexed transcripts via
$$s^* = \underset{s_i \in S}{\arg\max} \left(\frac{e_{\text{text}}(\tilde{q}) \cdot e_{\text{text}}(s_i)}{\|e_{\text{text}}(\tilde{q})\| \, \|e_{\text{text}}(s_i)\|} \right) \tag{5}$$
where S is the set of all subtitles in the video. Using the temporal window of the matched subtitle s^*, it returns the corresponding frames of the video clip v^* as o_t, enabling precise retrieval of specific spoken dialog or named entities that semantic visual embeddings miss.

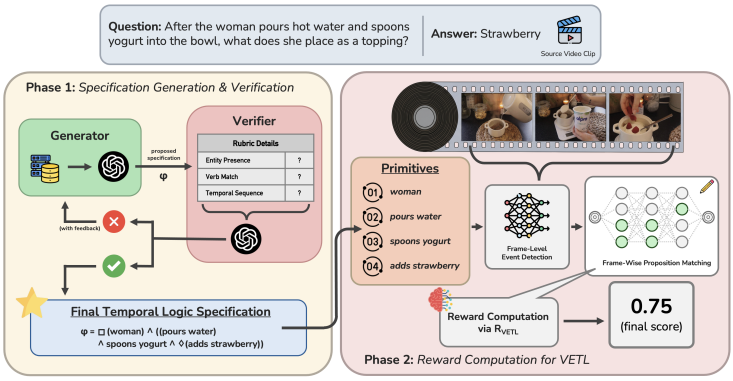

Fig. 2. VETL Rewards. We separately compile the question and its correct answer into a temporal logic specification in Phase 1. We utilize the extracted primitives from the temporal logic specification to verify whether the agent's retrieved video context would satisfy answering the question in Phase 2.

3. **Video Summary:** Invoked via the `<summary></summary>` tag, this tool gives a global view of the video. When the agent lacks priors for a targeted search, it can trigger this exploratory action. The environment skips the search index and returns 16 frames uniformly sampled throughout the duration of V, anchoring the agent and providing the temporal context needed to formulate more precise search queries later.

4.3 VETL: Visual Evidence via Temporal Logic

We introduce `VETL`, a neuro-symbolic framework that compiles a natural language question q and reference answer a into a TL specification, enabling dense rewards over τ.

Specification Generation. `VETL` derives reliable temporal-logic specifications via an iterative *generate–verify* loop (Algorithm 1). An LLM proposes a TL specification φ over a set of atomic primitives $\mathcal{P}$ (*e.g.*, object-centric predicates such as `car` or event-centric predicates such as `turn-left`), following prior work [58,55] [Line 3]. A judge model then scores φ under strict rubrics (in the Appendix) for entity coverage and temporal faithfulness [Line 4]. The judge also returns structured feedback to the generator, and the process repeats until the score is above the threshold κ set to 5 [Line 5].

Reward Composition. Once the formal specification φ is accepted, the agent's visual observations $(o_1, \ldots, o_T)$ are mapped to a dense reward. This reward incentivizes the agent to surface the specific visual evidence mandated by the specification. Let $\mathcal{P}_\varphi$ denote the finite set of unique visual primitives extracted

Algorithm 1 `VETL` Phase 1: Neuro-Symbolic Specification Generation

Require: Question q, reference answer a (optional), score threshold for acceptance κ, max rounds N, judge rubrics $\mathcal{R}$, LLM-based temporal specification generator GENERATOR, and LLM-based Judge JUDGE .
Ensure: Accepted TL specification φ.

```
fb ← ∅ {initialize structured feedback}
for r = 1 to N do
    φ ← GENERATOR(q, a, fb) {propose TL over primitives P}
    (score, fb) ← JUDGE(q, a, φ; R) {evaluate against rubrics}
    if score ≥ κ then
        break
    end if
end for
return φ
```

from φ (see Figure 2). We define a deterministic visual checker as a detection function using a Deep Learning Model (ViCLIP in our case) $D_p(o_t) \in \{0, 1\}$, which evaluates to 1 if primitive $p \in \mathcal{P}_\varphi$ is present in observation o_t, and 0 otherwise. The reward is the fraction of required primitives successfully observed across the entire retrieved video context:

$$R_{\text{VETL}}(\tau; \varphi) = \frac{1}{|\mathcal{P}_\varphi|} \sum_{p \in \mathcal{P}_\varphi} \max_{1 \le t \le T} D_p(o_t) \tag{6}$$

Notably, while φ is formalized in temporal logic [7,16], we use temporal operators only to disentangle semantics from their temporal relationships. Accordingly, we ignore the order in which evidence is retrieved, since the agent's search is non-chronological (*e.g.*, it may search for a later event in the prompt, first). The final reward is given by $R = \lambda_1 R_{\text{em}} + \lambda_2 R_{\text{VETL}}$. Here, $\lambda_1 + \lambda_2 = 1$, are set to 0.7 and 0.3 respectively, and R_{em} is the exact match reward.

5 Experiments

This section evaluates `VSeek` with a focus on verified-signal rewards for post-training and long-horizon video understanding with tool use.

5.1 Research Questions

We structure our experimental evaluation around four research questions:

1. **RQ1 (Reward Effectiveness):** How does a standard language-based search agent compare to one incentivized by our verified-signal reward?
2. **RQ2 (Long-Video Understanding):** Does the reward mechanism yield a significant advantage as video duration increases?
3. **RQ3 (Frame Efficiency):** What efficiency gains arise in computational overhead and average frames processed per query?
4. **RQ4 (Algorithmic Robustness):** Does adding a verified reward consistently improve performance across multiple post-training RL algorithms?

5.2 Experimental Setup

Model Zoo and Baselines. To evaluate the efficacy of the proposed framework, we benchmark against a diverse set of multimodal architectures categorized into static baselines and active agentic models. Our inference-only baselines include frontier models such as `GPT-5.2`, and `InternVL-3.5-4B`, along with Qwen3 variants such as `Qwen3-VL-4B-Thinking` and `Qwen3-VL-4B-Instruct`. For agentic evaluation, we compare our performance with VideoTree [69], whose captioning and QA frameworks are constrained on the `Qwen3-VL-4B-Instruct` model, and specialized variants of our framework, including Naive-RAG, `VSeek-GPT`, `VSeek-EM`, and `VSeek-VETL`. Naive-RAG retrieves the top 64 frames that have the highest similarity with respect to the original question, whereas `VSeek-GPT` performs the same natural language-based search with `GPT-5.2` as the base model. Both `VSeek-EM` and `VSeek-VETL` are post-trained specifically on top of `Qwen3-VL-4B-Thinking` models to leverage their native chain-of-thought capabilities for complex video search trajectories. `VSeek-EM` is post-trained with GRPO with the exact match reward, whereas `VSeek-VETL` incorporates the `VETL` reward defined in Equation 6.

Video Indexing and Retrieval. A core component of our architecture is the structured representation of raw video streams. We use ViCLIP to generate dense embeddings over segmented videos, indexing each video with fixed 8-second segments. This granularity lets the retriever efficiently align natural language queries with specific video clips and enables a hierarchical coarse-to-fine search strategy. Using the same 8-second indexing across all benchmarks ensures consistent event grounding for the agent's search actions.

Datasets and Benchmarks. Our evaluation covers both in-domain (ID) and out-of-domain (OOD) settings to probe whether the `VSeek` variants remain performant and reliable across multiple genres of questions. In-domain performance is measured on LongVideoBench [70], MLVU [79], and VideoMME [20], which require accurate retrieval over long videos. To assess zero-shot transfer, we also report results on CGBench-mini [10] and LVBench [66], challenging OOD benchmarks for testing the robustness of language-based search on unseen video distributions.

Evaluation Metrics. We evaluate performance using metrics based on success and efficiency. We measured the *Pass@1* and *Pass@4*, which capture the probability of generating one correct answer across 1 and 4 passes, respectively, and *Majority@4*, which assesses consensus accuracy across 4 independent samples. For efficiency, we report the average number of frames sampled per video, indicating how selectively the model searches for sparse visual evidence and the resulting computational cost for long-video understanding.

Implementation Details. Full hyperparameters and training prompts are in the Appendix. We train on 8 GH200s for 900 steps with a batch size of 32.

Table 1. Results on three ID (*) and two OOD benchmarks. `VSeek` is initialized from `Qwen3-VL-4B-Thinking`; for the post-trained variants `VSeek-EM` and `VSeek-VETL`, we report the *absolute score* (best in **bold** and second best in underline) and the *relative gain* (+/-%) with respect to direct inference on the base model (`Qwen3-VL-4B-Thinking`). Note that `VSeek-EM` is trained with an exact-match reward only, while `VSeek-VETL` uses `VETL` rewards. Additionally, we report the performance of the `GPT-5.2` variant of `VSeek` (`VSeek-GPT`).

Method	LongVideoBench*			Video-MME*			MLVU*			LVBench			CGBench-mini		
	p@1	p@4	m@4	p@1	p@4	m@4	p@1	p@4	m@4	p@1	p@4	m@4	p@1	p@4	m@4
Proprietary Models															
`GPT-5.2` (64 frames)	57.8	63.5	58.1	**67.7**	72.9	**67.9**	61.3	67.1	61.5	38.8	46.1	39.1	**36.7**	44.3	37.2
`VSeek-GPT`	54.4	**72.8**	57.2	61.2	**77.8**	65.3	56.3	75.4	59.9	**40.3**	**58.5**	40.9	34.4	**56.2**	**39.9**
Open Source Models															
`InternVL-3.5-4B`	38.9	60.4	42.9	53.1	76.1	54.7	41.8	69.1	43.4	32.6	57.7	33.2	21.5	49.2	22.3
`Qwen3-VL-4B-Thinking`	51.7	60.8	53.7	58.1	65.2	60.0	58.0	64.3	58.8	35.6	41.8	36.4	31.6	41.6	32.4
`Qwen3-VL-4B-Instruct`	54.4	59.2	54.9	63.7	68.6	64.3	61.9	66.2	62.4	38.7	47.7	38.6	34.7	43.5	34.7
`Qwen3-VL-8B-Thinking`	51.7	68.6	56.3	60.3	73.4	63.0	56.3	71.5	59.4	35.1	51.3	37.4	31.3	49.9	33.7
`Qwen3-VL-8B-Instruct`	54.5	59.8	54.9	63.7	68.6	64.3	61.9	66.2	62.4	38.7	47.7	38.6	34.7	43.5	34.7
Agentic Systems															
Naive RAG	49.8	60.9	49.8	55.0	63.3	56.5	55.6	64.0	57.5	37.1	46.1	38.7	30.4	41.2	32.2
VideoTree	45.4	47.6	45.1	56.3	58.0	56.6	46.1	48.2	46.8	34.0	36.4	34.2	29.6	31.9	30.0
`VSeek-EM`	60.5	71.1	62.1	59.3	71.2	60.1	66.7	75.9	67.3	40.0	54.9	**41.5**	33.2	52.3	34.5
	+8.8%	+10.3%	+8.4%	+1.2%	+6.0%	+0.1%	+8.7%	+11.6%	+8.5%	+4.4%	+13.1%	+5.1%	+1.6%	+10.7%	+2.1%
`VSeek-VETL`	**60.9**	70.7	**62.6**	61.4	72.0	62.8	**67.9**	**77.6**	**68.8**	38.8	53.0	39.7	32.3	51.1	33.3
	+9.2%	+9.9%	+8.9%	+3.3%	+6.8%	+2.7%	+9.9%	+13.3%	+10.0%	+3.2%	+11.2%	+3.3%	+0.7%	+9.5%	+0.9%

5.3 RQ1: Performance with the Verified-Signal Reward

Performance of Agentic Search. As shown in Table 1, the transition from passive perception to an active, natural language-driven search policy yields significant gains over existing video agent baselines. By utilizing a flexible language interface for evidence seeking, `VSeek` effectively pinpoints sparse evidence. `VSeek` variants consistently outperform other agentic systems like Naive RAG, with the most substantial improvements observed on ID datasets. While gains on OOD benchmarks are more modest, they remain positive. We expect that larger base models will further enhance the generalizability of this search policy, leading to more sophisticated cross-domain evidence acquisition.

Impact of `VETL` Rewards. The introduction of neuro-symbolic verified rewards from `VETL` provides a distinct advantage over exact match (EM) baselines by enforcing evidence coverage during video context construction. On in-domain evaluations, this verification signal drives a performance increase, specifically a 1.2% gain on MLVU for `VSeek-VETL` over `VSeek-EM`, as the model is explicitly rewarded for evidence seeking.

In contrast, we observe a modest degradation in Pass@1 scores compared to the agent trained with exact match reward on OOD evaluations. Error analysis suggests that this drop stems from the high specificity of learned search strings, which can become overly specialized to the in-distribution corpus. We believe that this is a "specialization bottleneck" rather than a fundamental flaw; a more diverse training corpus should allow the model to learn generalized search primitives and strategies that maintain high performance across unseen genres without sacrificing the rigor of the verification signal.

Table 2. RQ2 results: Performance (best in **bold** and second best in underline) of post-trained `VSeek` variants with gains (+/-%) reported relative to the `Qwen3-VL-4B-Thinking` base model. We include `VSeek-GPT` (utilizing `GPT-5.2` for reasoning) as a strong agentic baseline. Our post-trained models consistently outperform this variant. Additionally, we show that, particularly on longer videos (>10m on LongVideoBench and >1m in Video-MME and MLVU), the models post-trained with the verified-signal reward have better performance than the model variant post-trained with only the exact-match reward.

Method	LongVideoBench			Video-MME			MLVU		
	<1m	1–10m	10–60m	<1m	1–10m	10–60m	1–10m	10–60m	60m+
Baselines									
`Qwen3-VL-4B-Instruct`	62.5	56.8	48.9	68.1	**69.8**	57.7	63.4	59.9	51.2
`Qwen3-VL-4B-Thinking`	60.9	56.3	47.9	69.4	64.7	53.9	58.9	59.9	51.2
Naive RAG	58.1	55.5	46.2	56.8	60.9	52.2	56.8	60.8	61.2
VideoTree	40.2	44.4	40.8	61.0	58.5	53.8	44.7	55.3	40.5
Post-trained Variants									
`VSeek-EM`	69.1 +8.2%	**62.8** **+6.5%**	57.5 +9.6%	66.7 -2.7%	65.0 +0.3%	53.8 -0.1%	67.0 +8.1%	69.0 +9.1%	**62.2** **+11.0**
`VSeek-VETL`	**69.6** **+8.7%**	61.9 +5.6%	**58.9** **+11.0%**	68.6 -0.8%	66.6 +1.9%	57.7 +3.8%	**68.8** **+9.9%**	**69.9** **+10.0%**	**62.2** **+11.0%**
Proprietary Models									
`VSeek-GPT`	59.5	59.1	54.4	**69.5**	68.9	**60.6**	59.2	63.1	55.8

5.4 RQ2: Long Video Understanding with Tool Use

Our results show that both `VSeek-VETL` and `VSeek-EM` maintain substantially higher reasoning stability than baseline agentic systems as video length increases. While methods like VideoTree and Naive RAG preprocess and select scenes of interest separately from reasoning and answering, both `VSeek` variants interleave search directly with thinking and response generation. This integration enables more coherent answers on longer videos, as the model can dynamically adapt its search strategy based on intermediate reasoning and gathered evidence. Moreover, by awarding partial credit for verifiable intermediate steps via our neuro-symbolic pipeline, the verified reward provides a denser training signal than a binary success/failure metric, encouraging the model to first localize key events and rigorously verify them before synthesizing an answer.

In longer videos, `VSeek-VETL` outperforms `VSeek-EM` and VideoTree by providing more specific and accurate grounding. Qualitatively, we observe a shift in evidence-seeking behavior: `VSeek-EM` often issues generic queries which can retrieve irrelevant clips, whereas `VSeek-VETL` learns to produce highly specific keywords. Since these keywords do not directly match the atomic primitives from the `VETL` pipeline, rewards are devised based on the retrieved video context, to ensure that it is closely aligned with the query. Therefore, this search is enabled by the `VETL`-based reward, which trains the model to explore better search strings rather than repeatedly issuing unproductive search queries, a common failure of baseline agentic systems. Figure 3 shows how `VSeek-VETL` formulates more targeted queries to navigate a long video.

Because both `VSeek-VETL` and `VSeek-EM` are post-trained to answer questions with the search-based tool, both methods learn a precise internal model of the

Table 3. Efficiency Analysis (**RQ3**): Comparison of the average frames sampled across diverse video durations for three in-domain datasets. `VSeek-VETL` significantly reduces the visual footprint while achieving state-of-the-art performance.

Method	LongVideoBench				Video-MME				MLVU			
	<1m	1m-10m	10m-60m	**Avg.**	<1m	1m-10m	10m-60m	**Avg.**	1-10m	10m-60m	60m+	**Avg.**
Baselines												
`Qwen3-VL-4B-Instruct`	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0
`Qwen3-VL-4B-Thinking`	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0
Naive RAG	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0	64.0
VideoTree	13.6	206.7	384.3	232.6	26.5	111.3	256.6	168.8	213.7	323.5	402.4	240.7
VSeek Variants												
`VSeek-EM`	30.3	31.2	31.4	31.0	29.5	30.2	31.6	30.8	30.1	30.4	30.5	30.2
`VSeek-VETL`	31.9	31.9	32.1	32.0	31.8	31.9	32.0	31.9	31.9	31.6	31.3	31.8
Proprietary Models												
`VSeek-GPT`	17.6	27.1	26.9	24.5	20.4	25.7	32.9	28.5	23.9	23.5	30.1	28.9

video's temporal structure, allowing them to find sparse evidence in hour-long raw footage with highly targeted search actions. This post-training yields better reasoning capability, even outperforming `VSeek-GPT`, which uses `GPT-5.2` as the reasoning model to issue search queries and answer questions.

5.5 RQ3: Efficiency Gains

We evaluate `VSeek`'s computational efficiency via average frames processed per query. Standard VLMs and RAG baselines typically use a fixed 64-frame budget, whereas `VSeek` variants achieve better accuracy with under 20 frames on average. For VideoTree, which requires frame-level captions and samples these captions to adjudicate query relevancy for QA, we report the number of sampled captions. As shown in Table 3, `VSeek-VETL` samples slightly more frames than `VSeek-EM` (*e.g.*, 31.9 vs. 30.3 on LongVideoBench) due to the evidence-based reward where penalizing unverified claims encourages more thorough retrieval. This trade-off ensures that the retrieved context meets the temporal and causal requirements of the query, improving the quality of the reasoning. `VSeek-VETL` however does not outperform much larger proprietary systems like `VSeek-GPT`, which averages 24–29 frames on long-form content. Coupled with the degradation of the performance on these tasks, we observe that `GPT-5.2` often answers questions without complete information, resulting in the lower frame usage.

5.6 RQ4: Influence Across Post-Training RL Algorithms

Building on these initial gains, we investigate whether the benefits of verified rewards are consistent across diverse RL algorithms. We instantiate GRPO [57], Group Sequence Policy Optimization (GSPO) [78], and REINFORCE Leave-One-Out (RLOO) [2], using both exact match reward and our verified reward signal. We show that the `VETL` reward improves performance with GRPO and RLOO post-training algorithms, while showing degradation with GSPO. We provide these results in the Appendix.

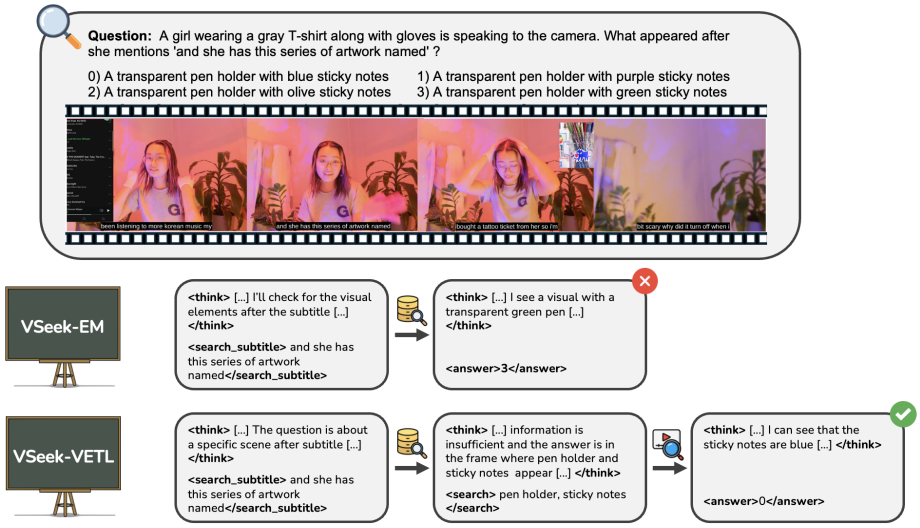

Fig. 3. Qualitative analysis. This example qualitatively highlights the advantage of `VSeek-VETL` over outcome-only based training. Here, both `VSeek-EM` and `VSeek-VETL` correctly decide that the first clip to retrieve must contain the desired subtitle. In contrast to `VSeek-EM`, which immediately hallucinates an answer, `VSeek-VETL` devises a more targeted search query to retrieve a set of segments corresponding to a pen holder to produce the correct answer.

6 Limitations and Future Work

Despite the performance gains achieved by `VSeek`, several promising avenues for future research remain.

1. **Non-Sequential Search Paradigms**: `VSeek` currently runs sequentially, which can cause out-of-order frame processing relative to the video timeline. Future work can adopt "fan-out" search strategies, similar to deep research frameworks, so that the agent can pursue multiple search strings in parallel.
2. **Advanced Formal Methods**: While `VETL` is a strong foundation, integrating more expressive temporal logic families can further refine the reward landscape. Especially, in conjunction with better agentic architectures, models can be rewarded better to capture more complex causal and temporal dependencies.
3. **Retriever and Indexing Bottlenecks**: Video evidence grounding is limited by the underlying retriever and indexing models. Future work can use interleaved search with stronger multimodal retrievers and explore generalization to diverse retrievers.

7 Conclusion

In this work, we make two contributions. First, we introduce `VSeek`, an agentic framework that leverages natural language-based search tools for long video question answering, therefore, moving beyond passive perception and uniform sampling. Then, to ensure that the retrieved context is relevant to the question, we further introduce `VETL`, a neuro-symbolic pipeline from which we derive dense, verified rewards for effective RL post-training to link high-level reasoning required for tool calls with low-level visual grounding.

Empirically, `VSeek` variants consistently outperform static baselines and match strong agentic models like `GPT-5.2`, especially on longer videos. We also show that our proposed reward `VETL` improves post-training a nd the final model's performance by ensuring that the agent's intermediate "thought-and-search" gathers relevant video evidence. We hope `VSeek` and `VETL` offer a solid foundation for future work on stronger video agents and post-training methods for long video understanding.

Acknowledgements

This material is based upon work supported in part by the Office of Naval Research (ONR) under Grant No. N00014-22-1-2254. Additionally, this work was supported by the Defense Advanced Research Projects Agency (DARPA) contract DARPA ANSR: RTX CW2231110. Approved for Public Release, Distribution Unlimited. We also like to thank the Texas Advanced Computing Center at The University of Texas at Austin for compute support.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., Hooker, S.: Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12248–12267 (2024)
3. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
7. Baier, C., Katoen, J.P.: Principles of Model Checking. The MIT Press (2008)
8. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? (2021), https://arxiv.org/abs/2102.05095
9. Chandrasegaran, K., Gupta, A., Hadzic, L.M., Kota, T., He, J., Eyzaguirre, C., Durante, Z., Li, M., Wu, J., Li, F.F.: Hourvideo: 1-hour video-language understanding. Advances in Neural Information Processing Systems **37**, 53168–53197 (2025)
10. Chen, G., Liu, Y., Huang, Y., He, Y., Pei, B., Xu, J., Wang, Y., Lu, T., Wang, L.: Cg-bench: Clue-grounded question answering benchmark for long video understanding. arXiv preprint arXiv:2412.12075 (2024)
11. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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. pp. 24185–24198 (2024)
12. Chen, Z., Yi, K., Li, Y., Ding, M., Torralba, A., Tenenbaum, J.B., Gan, C.: Comphy: Compositional physical reasoning of objects and events from videos. arXiv preprint arXiv:2205.01089 (2022)
13. Choi, M., Goel, H., Omama, M., Yang, Y., Shah, S., Chinchali, S.: Towards neuro-symbolic video understanding. In: European Conference on Computer Vision. pp. 220–236. Springer (2024)
14. Choi, M., Sharan, S., Goel, H., Shah, S., Chinchali, S.: We'll fix it in post: Improving text-to-video generation with neuro-symbolic feedback. arXiv preprint arXiv:2504.17180 (2025)
15. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
16. Emerson, E.A.: Temporal and modal logic. In: Handbook of Theoretical Computer Science, Volume B: Formal Models and Sematics (1991), https://api.semanticscholar.org/CorpusID:6062082
17. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
18. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1933–1941. IEEE Computer Society, Las Vegas, NV, USA (2016)
19. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
20. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
21. Goel, H., Udathu, A., Jabireddy, S., Kalkar, P., Parulekar, A.: Sˆ 3-r1: Learning to retrieve and answer step-by-step with synthetic data. arXiv preprint arXiv:2605.01248 (2026)

22. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
23. Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., Chen, G., Bi, X., Wu, Y., Li, Y., et al.: Deepseek-coder: when the large language model meets programming–the rise of code intelligence. arXiv preprint arXiv:2401.14196 (2024)
24. Hasanbeig, M., Kantaros, Y., Abate, A., Kroening, D., Pappas, G.J., Lee, I.: Reinforcement learning for temporal logic control synthesis with probabilistic satisfaction guarantees. In: 2019 IEEE 58th conference on decision and control (CDC). pp. 5338–5343. IEEE (2019)
25. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13504–13514 (2024)
26. He, Z., Qu, X., Li, Y., Huang, S., Liu, D., Cheng, Y.: Framethinker: Learning to think with long videos via multi-turn frame spotlighting. arXiv preprint arXiv:2509.24304 (2025)
27. Hong, K., Troynikov, A., Huber, J.: Context rot: How increasing input tokens impacts llm performance. URL https://www.trychroma.com/research/context-rot, retrieved October **20**, 2025 (2025)
28. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: European conference on computer vision. pp. 709–727. Springer (2020)
29. Islam, M.M., Nagarajan, T., Wang, H., Bertasius, G., Torresani, L.: Bimba: Selective-scan compression for long-range video question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29096–29107 (2025)
30. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2758–2766 (2017)
31. Jeoung, S., Huybrechts, G., Ganesh, B., Galstyan, A., Bodapati, S.: Adaptive video understanding agent: Enhancing efficiency with dynamic frame sampling and feedback-driven reasoning. arXiv preprint arXiv:2410.20252 (2024)
32. Jha, S., Raman, V., Sadigh, D., Seshia, S.A.: Safe autonomy under perception uncertainty using chance-constrained temporal logic. Journal of Automated Reasoning **60**, 43–62 (2018)
33. Jiang, F., Yuan, J., Tsaftaris, S.A., Katsaggelos, A.K.: Anomalous video event detection using spatiotemporal context. Comput. Vis. Image Underst. **115**(3), 323–333 (2011). https://doi.org/10.1016/J.CVIU.2010.10.008, https://doi.org/10.1016/j.cviu.2010.10.008
34. Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., Han, J.: Search-r1: Training llms to reason and leverage search engines with reinforcement learning. arXiv preprint arXiv:2503.09516 (2025)
35. Kress-Gazit, H., Fainekos, G.E., Pappas, G.J.: Temporal-logic-based reactive mission and motion planning. IEEE Transactions on Robotics **25**(6), 1370–1381 (2009). https://doi.org/10.1109/TRO.2009.2030225
36. Kroshchanka, A., Golovko, V., Mikhno, E., Kovalev, M., Zahariev, V., Zagorskij, A.: A neural-symbolic approach to computer vision. In: Golenkov, V., Krasnoproshin, V., Golovko, V., Shunkevich, D. (eds.) Open Semantic Technologies for Intelligent Systems. pp. 282–309. Springer International Publishing, Cham (2022)

37. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
38. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
39. Li, N., Chang, F., Liu, C.: Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. Neurocomputing **490**, 482–494 (2022)
40. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
41. Liang, S., Shah, S., Zhou, C., Sharan, S., Goel, H., Sanyal, A., Chinchali, S., Datta, G.: Le-neus: Latency-efficient neuro-symbolic video understanding via adaptive temporal verification. arXiv preprint arXiv:2602.23553 (2026)
42. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
43. Medioni, G., Cohen, I., Bremond, F., Hongeng, S., Nevatia, R.: Event detection and analysis from video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(8), 873–889 (2001). https://doi.org/10.1109/34.946990
44. Mehdipour, N., Althoff, M., Tebbens, R.D., Belta, C.: Formal methods to comply with rules of the road in autonomous driving: State of the art and grand challenges. Automatica **152**, 110692 (2023). https://doi.org/https://doi.org/10.1016/j.automatica.2022.110692, https://www.sciencedirect.com/science/article/pii/S0005109822005568
45. Munir, M., Goel, H., Wei, X., Choi, M., Shah, S., Bhardwaj, K., Whatmough, P., Chinchali, S., Marculescu, R.: Objectalign: Neuro-symbolic object consistency verification and correction. arXiv preprint arXiv:2511.18701 (2025)
46. Murphy, W., Holzer, N., Koenig, N., Cui, L., Rothkopf, R., Qiao, F., Santolucito, M.: Guiding llm temporal logic generation with explicit separation of data and control. arXiv preprint arXiv:2406.07400 (2024)
47. Pan, J., Zhang, Q., Zhang, R., Lu, M., Wan, X., Zhang, Y., Liu, C., She, Q.: Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. arXiv preprint arXiv:2511.05489 (2025)
48. Pnueli, A.: The temporal logic of programs. 18th Annual Symposium on Foundations of Computer Science (FOCS) pp. 46–57 (1977)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
50. Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., et al.: Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950 (2023)
51. Sarkar, S., Lore, K.G., Sarkar, S.: Early detection of combustion instability by neural-symbolic analysis on hi-speed video. In: CoCo@ NIPS (2015)
52. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach

themselves to use tools. Advances in neural information processing systems **36**, 68539–68551 (2023)
53. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
54. Shah, S., Goel, H., Narasimhan, S.S., Choi, M., Sharan, S., Akcin, O., Chinchali, S.: A challenge to build neuro-symbolic video agents. arXiv preprint arXiv:2505.13851 (2025)
55. Shah, S., Sharan, S.P., Goel, H., Choi, M., Munir, M., Pasula, M., Marculescu, R., Chinchali, S.: Neus-qa: Grounding long-form video understanding in temporal logic and neuro-symbolic reasoning. Proceedings of the AAAI Conference on Artificial Intelligence **40**(11), 8805–8813 (Mar 2026). https://doi.org/10.1609/aaai.v40i11.37834, https://ojs.aaai.org/index.php/AAAI/article/view/37834
56. Shao, Z., Luo, Y., Lu, C., Ren, Z., Hu, J., Ye, T., Gou, Z., Ma, S., Zhang, X.: Deepseekmath-v2: Towards self-verifiable mathematical reasoning. arXiv preprint arXiv:2511.22570 (2025)
57. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
58. Sharan, S., Choi, M., Shah, S., Goel, H., Omama, M., Chinchali, S.: Neuro-symbolic evaluation of text-to-video models using formal verification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8395–8405 (2025)
59. Song, E., Chai, W., Ye, T., Hwang, J.N., Li, X., Wang, G.: Moviechat+: Question-aware sparse memory for long video question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
60. Song, H., Jiang, J., Min, Y., Chen, J., Chen, Z., Zhao, W.X., Fang, L., Wen, J.R.: R1-searcher: Incentivizing the search capability in llms via reinforcement learning. arXiv preprint arXiv:2503.05592 (2025)
61. Tan, X., Luo, Y., Ye, Y., Liu, F., Cai, Z.: Allvb: All-in-one long video understanding benchmark (2025), https://arxiv.org/abs/2503.07298
62. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)
63. Team, C., Zhao, H., Hui, J., Howland, J., Nguyen, N., Zuo, S., Hu, A., Choquette-Choo, C.A., Shen, J., Kelley, J., et al.: Codegemma: Open code models based on gemma. arXiv preprint arXiv:2406.11409 (2024)
64. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4975–4986 (2021)
65. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5552–5561 (2019)
66. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22958–22967 (2025)
67. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)

68. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023)
69. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3272–3283 (2025)
70. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
71. Xue, Z.S., Luo, R., Grauman, K.: Seeing the arrow of time in large multimodal models. Advances in Neural Information Processing Systems **38**, 90925–90955 (2026)
72. Yang, Y., Gaglione, J.R., Chinchali, S., Topcu, U.: Specification-driven video search via foundation models and formal verification. arXiv preprint arXiv:2309.10171 (2023)
73. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
74. Ye, J., Wang, Z., Sun, H., Chandrasegaran, K., Durante, Z., Eyzaguirre, C., Bisk, Y., Niebles, J.C., Adeli, E., Fei-Fei, L., et al.: Re-thinking temporal search for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8579–8591 (2025)
75. Ye, Y., Zhao, Z., Li, Y., Chen, L., Xiao, J., Zhuang, Y.: Video question answering via attribute-augmented attention network learning. In: Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 829–832 (2017)
76. Yu, D., Yang, B., Wei, Q., Li, A., Pan, S.: A probabilistic graphical model based on neural-symbolic reasoning for visual relationship detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10599–10608 (2022). https://doi.org/10.1109/CVPR52688.2022.01035
77. Yuan, H., Liu, Z., Zhou, J., Qian, H., Shu, Y., Sebe, N., Wen, J.R., Dou, Z.: Videoexplorer: Think with videos for agentic long-video understanding. arXiv preprint arXiv:2506.10821 (2025)
78. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al.: Group sequence policy optimization. arXiv preprint arXiv:2507.18071 (2025)
79. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025)
80. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852 (2023)
81. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4995–5004 (2016)