Towards Sparse Video Understanding and Reasoning

Chenwei Xu[1] Zhen Ye[2] Shang Wu[1] Weijian Li[1] Zihan Wang[1] Zhuofan Xia[1]
Lie Lu[3] Pranav Maneriker[3] Fan Du[3] Manling Li[1] Han Liu[1]
[1] Northwestern University [2] Johns Hopkins University [3] Dolby Laboratories

Abstract

We present REVISE *(Reasoning with Video Sparsity), a multi-round agent for video question answering (VQA). Instead of uniformly sampling frames,* REVISE *selects a small set of informative frames, maintains a summary-as-state across rounds, and stops early when confident. It supports proprietary vision-language models (VLMs) in a "plug-and-play" setting and enables reinforcement fine-tuning for open-source models. For fine-tuning, we introduce EAGER (Evidence-Adjusted Gain for Efficient Reasoning), an annotation-free reward with three terms: (1) Confidence gain: after new frames are added, we reward the increase in the log-odds gap between the correct option and the strongest alternative; (2) Summary sufficiency: at answer time we re-ask using only the last committed summary and reward success; (3) Correct-and-early stop: answering correctly within a small turn budget is rewarded. Across multiple VQA benchmarks,* REVISE *improves accuracy while reducing frames, rounds, and prompt tokens, demonstrating practical sparse video reasoning.*

1. Introduction

Video understanding is challenging because video data are high-dimensional, temporally redundant, and semantically intricate. Recent progress in large language models (LLMs) has accelerated video understanding research. Many recent works [1, 3, 27, 50, 59, 63] have made steady progress toward steering vision-language models (VLMs) to address these challenges. These approaches exploit LLMs' long-context reasoning capabilities to improve question answering over video [42, 58, 71]. In practice, videos are typically represented as a sequence of uniformly sampled frames. There are two major paradigms for integrating LLMs with those frames. The first involves using video captioning models to convert frames into textual descriptions, then utilizing LLMs [20, 59, 63] to perform analysis in textual space. However, this method may overlook fine-grained visual details inherently present in individual frames. To mitigate this limitation, the second directly integrates visual inputs into LLMs via pretrained vision encoders [76], forming vision-language models [3, 5, 27, 36]. However, these methods select frames uniformly, which still has two limitations, as shown in Wang et al. [63]: **(L1) Information Overload**: Long videos inherently exhibit substantial temporal redundancy. Too many redundant frames can overwhelm LLMs and hinder both reasoning and efficiency. **(L2) Insufficient Key Information Awareness**: Video content is hierarchically and temporally structured; without identifying semantically salient frames across scales, LLMs often miss critical cues for accurate reasoning. Both limitations stem from semantic sparsity. Only a small number of frames are relevant to a given question.

To address these challenges, we introduce REVISE (Reasoning with Video Sparsity), a multi-round agent for video question answering (VQA). Rather than processing a fixed set of uniformly sampled frames, REVISE iteratively (i) selects a small batch of frames most likely to reduce uncertainty about the answer, (ii) updates a concise summary of previous-round conversations (summary-as-state) to reduce information overload, and (iii) stops early once the accumulated evidence is sufficient to answer. Conceptually, our summary-as-state is inspired by the hidden state in recurrent neural networks (e.g., LSTMs) [16], as in Figure 1: it is a compact, continually updated memory that carries forward only task-critical information, informs what to "attend to next", and regularizes reasoning to remain faithful to accumulated evidence. This recurrent, stateful formulation reduces information overload **(L1)** by concentrating the visual context into non-redundant text-based evidence. Hence, this stateful design counters semantic sparsity by progressively accumulating a compact set of query-supporting frames in the persistent summary.

To improve key information awareness in VLMs **(L2)**, REVISE uses a structured summary-as-state that explicitly tracks what has been observed, how beliefs are updated, what remains uncertain, and why additional evidence is needed. This persistent state is compact, updated each round, and is the only information carried across turns, enabling the agent to recall prior evidence and request frames that target specific informational gaps. We show a detailed example in Figure 2. Instead of reprocessing long conversation histories or large

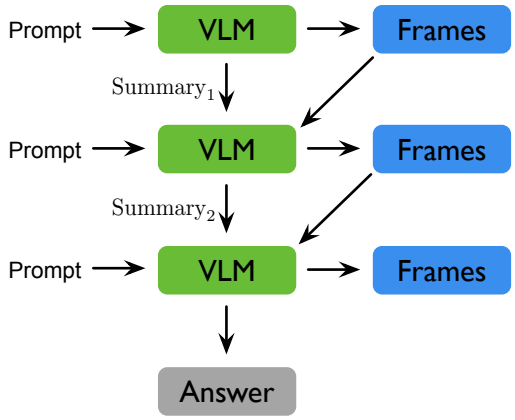

Figure 1. **Summary-as-State.** REVISE operates analogously to a recurrent neural network: it maintains a state that propagates information from previous turns to the VLM.

sets of previously seen frames, the agent continually updates this compact summary with only the verified, task-relevant evidence accumulated so far. By grounding each decision on this evolving state rather than raw past outputs, REVISE achieves coherent, uncertainty-aware multi-round reasoning while avoiding redundant processing, reducing token usage, and keeping the reasoning focused on the information that matters. REVISE operates in a fully "plug-and-play" manner: it wraps around any existing VLM without modifying its parameters or inference pipeline. Moreover, for open-source VLMs, REVISE can be paired with verifier-guided reinforcement fine-tuning [12, 62], further strengthening the model's ability to identify question-critical content, reduce frame redundancy, and execute more accurate and efficient video reasoning.

Our contribution is twofold: (I) Methodologically, we propose REVISE, a novel framework for question-aware video understanding. REVISE addresses the limitations in **(L1)** and **(L2)** by interactively selecting informative frames through a multi-round reasoning process. Its behavior is governed by two key components: (1) a multi-round conversation module that progressively gathers evidence and writes an evolving "summary-as-state" capturing the verified information across turns; and (2) a structured summary-as-state that records observations, belief updates, uncertainties, and selection rationale, and is the only information carried across rounds for stable, uncertainty-aware reasoning. REVISE is lightweight and modular: it wraps around any existing VLM in a plug-and-play manner without parameter updates. Further, when paired with verifier-guided reinforcement fine-tuning, REVISE strengthens open-source VLMs' ability to locate question-critical frames and reason over long videos more efficiently. (II) Experimentally, we conduct extensive evaluations across diverse video understanding benchmarks, spanning short clips to hour-long videos. In the plug-and-play setting, REVISE improves efficiency while matching or exceeding strong proprietary VLM baselines. With reinforcement fine-tuning, REVISE further boosts the performance of open-source VLMs, achieving higher accuracy with significantly fewer frames, fewer rounds, and fewer prompt tokens. Together, these results demonstrate that REVISE delivers practical and scalable sparse video reasoning.

2. Related Works

VLMs for Video Understanding. Several recent approaches extend image-centric vision-language models [2, 34, 57, 71] to video by treating videos as sequences of frames. Many methods [7, 14, 19, 23, 29–31, 33, 36, 41, 43, 61, 65, 73, 78] train vision-language models by connecting a vision encoder to a large language model through a lightweight adapter. Training-free methods [10, 20, 22, 52, 56, 59, 60, 63] integrate captioners with LLMs to support video understanding without full retraining. These pipelines also support interactive selection and open-ended relational reasoning for video QA [38, 59]. Specifically, LLoVi [77] generates short-term video captions using a visual encoder and prompts an LLM to summarize and answer user queries. VideoAgent [59] introduces an LLM-driven multi-round frame selection strategy based on captioned content. Recent VLMs such as LLaVA [27, 34, 36] and QwenVL [2, 3, 55] integrate stronger vision-language reasoning capabilities and reduce reliance on external captioners. REVISE combines agent-based search from VideoAgent with the direct visual reasoning abilities of modern VLMs.

Adaptive Frame Selection for Video LLMs. A growing body of work addresses the challenge of selecting informative frames from videos to improve VLM efficiency and accuracy. *Training-free approaches* avoid any parameter updates: MDP3 [51] formulates list-wise frame selection as a Markov decision process solved without training; Q-Frame [81] uses CLIP-based text-image matching with Gumbel-Max sampling for query-aware selection; FRAG [18] asks the model itself to score each frame's relevance; and FOCUS [82] casts keyframe selection as a combinatorial pure-exploration bandit problem. *Learned selectors* train lightweight policies: Frame-Voyager [74] learns to query task-relevant frames for video LLMs; Flexible Frame Selection [6] proposes a differentiable top-k selection operation trained end-to-end; Adaptive Keyframe Sampling [54] learns to sample keyframes tailored to long videos; M-LLM [17] leverages multimodal LLMs to guide frame selection; and K-frames [69] performs scene-driven any-k keyframe selection. *RL-based methods* optimize frame selection policies via reinforcement learning: TSPO [53] learns a temporal sampling policy with policy optimization; ReFoCUS [26] uses reinforcement-guided frame optimization for contextual understanding; FrameMind [11] enables frame-interleaved reasoning via GRPO; and FrameThinker [15] combines SFT with GRPO for multi-turn frame spotlighting. *Iterative agent-based approaches* perform multi-round frame selection guided by reasoning:

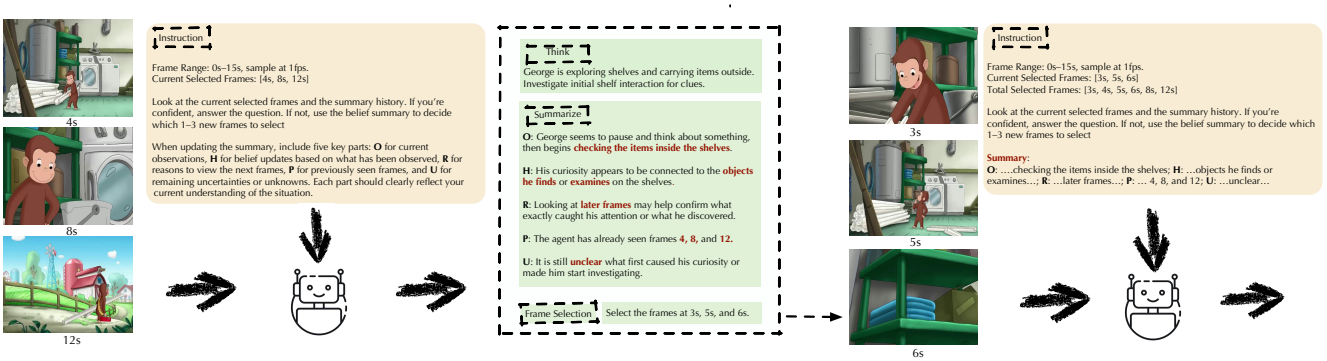

Figure 2. **Overview of REVISE: multi-round reasoning and adaptive frame selection.** Given an initial set of frames and a question, the VLM agent infers the video context to update the summary and selects relevant frames based on its reasoning. In the next round, the agent reasons over the selected frames and the updated summary to generate the final answer.

Active Video Perception [64] employs a plan-observe-reflect loop; A.I.R. [84] uses adaptive, iterative, reasoning-based selection; and VideoBrain [83] deploys dual complementary agents with behavior-aware rewards. Unlike prior work that typically performs a single-pass selection, REVISE maintains a persistent *summary-as-state* across rounds and couples it with EAGER reward for RL fine-tuning, enabling question-aware, iterative frame selection that remains compatible with any VLM in a plug-and-play manner.

VLM Reasoning and Multi-round Conversation. Recent progress in vision-language models (VLMs) has shifted from single-turn processing to interactive, reasoning-centric systems capable of sustaining multi-round dialog. Early efforts [67, 68, 75] enhanced interaction through prompt engineering and external APIs, bypassing the need for fully end-to-end architectures. More recent approaches incorporate explicit reasoning, enabling VLMs to infer answers based on implied visual information. For instance, DetGPT [46] performs object detection through high-level instructions rather than predefined class labels. GPT4RoI [80] uses spatial boxes to focus attention on specific regions, improving alignment between vision and language. Similarly, LISA [25] augments the mask decoder in SAM [21] with a learned embedding prompt, enabling high-level visual reasoning when paired with LLaVA [34]. Complementing these architectural advances, reinforcement learning techniques have emerged as effective tools for enhancing multi-step reasoning. PPO-based ReFT [39] rewards correct chains of thought, while DeepSeek-R1 [12] introduces step-wise rewards for logical soundness. DeepSeek-R1-Zero [12] demonstrates that outcome-only rewards can suffice when reasoning is self-verifiable. RAGEN [62] further shows that intermediate rewards are essential in preventing dialog agents from adopting shallow heuristics. Finally, self-revision models like Draft-Edit [24], S2R [40], and token-level reward "dancing" [32] underscore the value of iterative feedback for improving reasoning depth. Overall, these studies show that feedback and revision can drive multi-turn reasoning, ranging from reward shaping to simplified feedback signals [24, 32, 37, 40, 62]. Together, these developments point toward a new generation of VLMs that not only ground visual input accurately but also engage in self-correcting, multi-round reasoning to produce coherent and reliable answers.

3. Methodology

As illustrated in Figure 3, REVISE couples multi-round interaction with an explicit, structured summary-as-state. It targets two limitations of current VLMs in video understanding [63]: **(L1)** information overload and **(L2)** insufficient key-information awareness. We cast video understanding as an iterative, question-aware frame-selection problem that admits only frames most likely to support the query. REVISE mitigates **(L1)** via budgeted multi-round frame selection and improves **(L2)** by tracking observations, belief updates, uncertainties, and selection rationale in a persistent summary.

Problem Formulation. Given a video $V = \{x_i\}_{i=0}^{L-1}$ consisting of L frames, with the frame at time i denoted by x_i, and a user prompt p, the goal is to produce an answer a with a VLM agent π_θ while respecting a maximum context budget K. Let $c(x_i)$ denote the model-specific visual token cost of frame x_i, and $C(F) = \sum_{x \in F} c(x)$ the cost of a subset $F \subseteq V$. Instead of processing all frames (which may violate K), we construct the visual context *iteratively* and maintain a compact, persistent *"summary-as-state"*.

We model the interaction over at most T rounds. Let $S_t = \bigcup_{j=1}^{t} F_j$ be the set of all frames admitted up to round t (with $S_0 = \varnothing$). Let p_t denote the prompt at round t, constructed from the original prompt p, the previous summary z_{t-1}, and the shown frames F_t (with timestamps and basic video meta). The agent maintains a summary state:

$$z_t \;=\; (P_t,\, O_t,\, H_t,\, U_t,\, R_t), \tag{3.1}$$

where P_t (previously seen) summarizes what has already been inspected, O_t (observations) records the currently observed evidence, H_t (belief updates) captures how those

observations update the hypothesis (without outputting the final answer letter), U_t (uncertainties) enumerates remaining unknowns, and R_t (reasons) states what evidence to look for next (or that the question is answered). The order is fixed as $P \to O \to H \to U \to R$. Operationally, z_t is written explicitly in the `<summary>` field and is the *only* state carried across rounds. *Each turn conditions on all previous states*: the agent has access to the set of committed summaries $\{z_j\}_{j=0}^{t-1}$; since z_{t-1} is defined to be *cumulative*, conditioning on $\{z_j\}_{j<t}$ is equivalent to conditioning on the latest z_{t-1} alone.

At round t, given (p_t, z_{t-1}, F_t), the agent takes a single action

$$a_t \in \big\{\textsc{Select}(Q_t),\ \textsc{Answer}(y_t)\big\}, \tag{3.2}$$

where $Q_t \subseteq \{0, \ldots, L-1\} \setminus S_{t-1}$ is a small set of requested frame indices (0-based). If $a_t = \textsc{Select}(Q_t)$, the environment retrieves $F_{t+1} = \{x_i : i \in Q_t\}$, updates $S_{t+1} = S_t \cup F_{t+1}$, and the agent commits the next summary

$$z_t \ = \ f_\theta\big(z_{t-1},\, p_t,\, F_{t+1}\big) \ = \ \big(P_t,\, O_t,\, H_t,\, U_t,\, R_t\big), \tag{3.3}$$

with R_t explicitly guiding the proposal of the next Q_{t+1}. If $a_t = \textsc{Answer}(y_t)$, the process stops with $a = y_t$ at stopping time $\tau \le T$. At all times we enforce the token budget $C(S_t) + |p_t| \le K$.

Our objective is to choose the selection sequence $\{Q_t\}_{t=1}^{\tau-1}$ and stopping time τ (implicitly via the agent policy) to maximize task performance under budget:

$$\max_{\{Q_t\},\, \tau \le T} \ \mathcal{R}\Big(\pi_\theta\big(p,\, z_{\tau-1},\, S_{\tau-1}\big)\Big) \quad \text{s.t. } C(S_{\tau-1}) + |p_\tau| \le K. \tag{3.4}$$

This sequential formulation replaces single-shot maximum coverage with a summary-driven, question-aware selection process: the agent repeatedly (i) reads a few frames, (ii) updates $z_t = (P, O, H, U, R)$ so it cumulatively encodes all previous states, and (iii) decides (based on R_t and U_t) what to view next or when to answer. We detail how ReVISE instantiates the multi-round framework and how the summary-as-state enables efficient, query-focused reasoning below.

3.1. ReVISE Building Components

The core concept of ReVISE is to admit only frames that are relevant to the user's request while maintaining a compact state that carries verified evidence across rounds. Accordingly, ReVISE comprises three modules as in Figure 3: (i) a multi-round controller, (ii) a structured response format that externalizes the summary state, and (iii) a persistent summary-as-state. The multi-round controller adaptively selects frames and decides when to answer, while the summary-as-state is the sole memory that accumulates and conditions future decisions.

Multi-Round Controller. The agent interacts for at most T rounds. Let F_t be the frames shown at round t and $S_t = \bigcup_{j=1}^{t} F_j$ the union of all admitted frames ($S_0 = \varnothing$). Round t=1 starts from a small, uniformly sampled set F_1 and the initial prompt p (the prompt includes timestamps and basic video meta). At each subsequent round $t \ge 2$, given (p_t, z_{t-1}, F_t), the agent produces a single action:

$$a_t \in \{\textsc{Select}(Q_t), \textsc{Answer}(y_t)\}, \tag{3.5}$$

where $Q_t \subseteq \{0, \ldots, L-1\} \setminus S_{t-1}$ requests a few new indices (0-based). If $a_t = \textsc{Select}(Q_t)$, the environment fetches $F_{t+1} = \{x_i : i \in Q_t\}$, updates $S_{t+1} = S_t \cup F_{t+1}$, and the agent commits the next summary z_t. If $a_t = \textsc{Answer}(y_t)$, the process terminates. We enforce the token budget $C(S_t) + |p_t| \le K$ throughout. *Each turn conditions on all previous states* via the cumulative summary: by construction, z_{t-1} subsumes $\{z_0, \ldots, z_{t-2}\}$, so conditioning on z_{t-1} is equivalent to conditioning on the entire state history with constant memory cost. For notational clarity, the action is produced as:

$$a_t = \pi_\theta(p_t, z_{t-1}, F_t). \tag{3.6}$$

Structured Response Format. To make decisions transparent and to improve the quality of summary state, each response follows one of two formats:

SELECT:
`<summary>...</summary>` `<frames>...</frames>`.

ANSWER:
`<summary>...</summary>` `<answer>...</answer>`.

Here, `<summary>` is the *only* information persisted to the next round. This design keeps the prompt compact, improves interpretability, and guides query-aware selection and early stopping in subsequent rounds.

Summary-as-State. The persistent state is written explicitly in `<summary>` as

$$z_t = (P_t, O_t, H_t, U_t, R_t), \tag{3.7}$$

where P_t (Previously seen) summarizes what has already been inspected, O_t (Observations) records what was just seen, H_t (belief Hypotheses/updates) captures how those observations change the current hypothesis, U_t (Uncertainties) enumerates remaining unknowns, and R_t (Reasons) justifies which frames to view next (or indicates the question is answered). By carrying forward only z_t rather than raw histories, ReVISE (i) avoids information overload by not re-admitting redundant context and (ii) improves key-information awareness by explicitly tracking what has been observed (O), how beliefs change (H), what remains uncertain (U), and what to request next and why (R). The

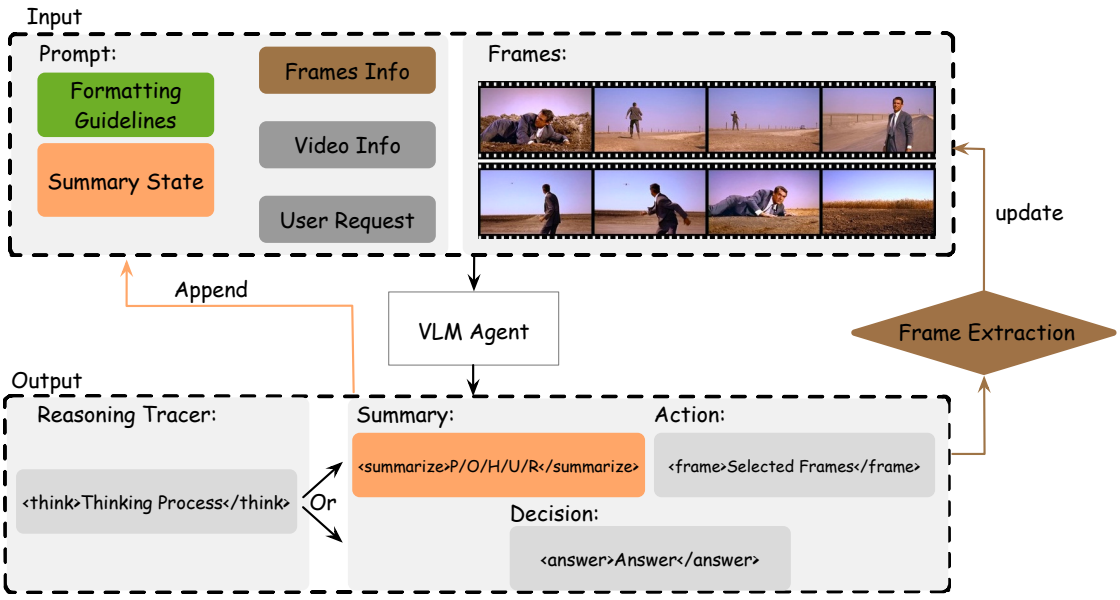

Figure 3. **REVISE**. REVISE consists of three components: multi-round conversation, a structured output protocol, and a summary-as-state. Each round, the VLM agent receives (i) the entire conversation history, (ii) the current prompt, and (iii) the chosen video frames, annotated with their timestamps and the video's total frame count. In the first round, a formatting guideline is also provided. The VLM outputs `<summary>` plus either `<frames>` (request) or `<answer>` (final), with the summary carrying the persistent state across rounds. REVISE then extracts the new frames and starts the next round with updated prompt and history. The conversation ends when the VLM produces a valid answer or the maximum number of rounds is reached.

R_t component directly informs the next proposal Q_{t+1}, and the stability of H_t/U_t provides a natural signal for when to answer. At each round, the policy performs multi-step reasoning to decide what to view next, but we do not expose a free-form chain-of-thought. Instead, the (H_t, U_t, R_t) fields in `<summary>` record a compact thinking process: hypothesis updates, remaining uncertainties, and reasons for the next action.

3.2. Plug-and-Play

REVISE treats any proprietary vision-language model as a frozen black-box module and communicates with it solely through its public inference interface (e.g., an API). All operations, including multi-round conversation, adaptive frame selection, structured summary updates, and validity enforcement, run externally within the framework, so no parameter updates are required. This design lets REVISE plug into closed-source systems as is, automatically orchestrating iterative multi-round interactions while leaving the model's original weights and vision-processing capabilities unchanged.

3.3. Reinforcement Fine-Tuning

Following Wang et al. [62], we cast the REVISE multi-round interaction as a finite-horizon MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma \rangle$ with horizon T. At round t, the state is

$$s_t = (p_t,\ z_{t-1},\ S_{t-1}), \tag{3.8}$$

where p_t is the current prompt (with formatting/meta), z_{t-1} is the cumulative summary-as-state committed in the previous round, and $S_{t-1} = \bigcup_{j=1}^{t-1} F_j$ is the set of admitted frames so far. The action space is

$$a_t \in \{\textsc{Select}(Q_t),\ \textsc{Answer}(y_t)\}, \tag{3.9}$$

where $Q_t \subseteq \{0, \ldots, L-1\} \setminus S_{t-1}$ requests new frame indices (0-based) and y_t is an answer. The transition updates (S_t, z_t): if $a_t = \textsc{Select}(Q_t)$, the environment returns $F_{t+1} = \{x_i : i \in Q_t\}$, sets $S_{t+1} = S_t \cup F_{t+1}$, and the agent commits the next summary $z_t = f_\theta(z_{t-1}, p_t, F_{t+1})$; if $a_t = \textsc{Answer}(y_t)$, the episode terminates at $\tau \le T$ with answer $a = y_t$. We optimize the expected return

$$J(\theta) = \mathbb{E}_{H \sim \pi_\theta}\left[\sum_{t=1}^{\tau} \gamma^{t-1} r_t\right], \tag{3.10}$$

and decompose π_θ into token-level likelihoods to remain compatible with autoregressive VLMs.

Reward Design: EAGER. We design a dense, annotation-free reward that aligns the policy with efficient, summary-driven reasoning. Let $\mathcal{Y}$ be the answer set (e.g., MCQ options) and $y^\star \in \mathcal{Y}$ the correct label. Define a (temperature-calibrated) log-odds margin under the current context

$$m_t = \log p_\theta\big(y^\star \mid p_t, z_{t-1}, S_t\big) - \max_{y \neq y^\star} \log p_\theta\big(y \mid p_t, z_{t-1}, S_t\big), \tag{3.11}$$

computed at each decision state (before taking action a_t). EAGER comprises three parts:

(i) Confidence gain. Reward only evidence that *truly* increases certainty after adding new frames:

$$r_t^{\text{conf}} = \big[m_{t+1} - m_t\big]_+ \quad \text{(applies on SELECT)}. \tag{3.12}$$

(ii) Summary sufficiency. At answer time, re-ask using *summary-only* to encourage faithful, compact state:

$$r_t^{\text{sum}} = \mathbb{1}\Big[\arg\max_{y\in\mathcal{Y}} p_\theta\big(y \mid p_\tau, z_\tau\big) = y^\star\Big] \quad \text{(applies on ANSWER)}. \tag{3.13}$$

(iii) Correct-and-early stop. We reward answering correctly within a small turn budget T_{stop}:

$$r_t^{\text{stop}} = \begin{cases} 1 + \beta\,\big[T_{\text{stop}} - \tau\big]_+, & a_\tau = \text{ANSWER}(y^\star) \text{ and } \tau \le T_{\text{stop}}, \\ 0, & \text{otherwise}, \end{cases} \tag{3.14}$$

applied at the final step. We also include a small structural bonus $r_t^{\text{format}} = \alpha \cdot \mathbb{1}[\text{valid format}]$ for emitting valid tags (`<summary>` plus `<frames>` or `<answer>`, with no extra text).

The per-step reward is

$$r_t \;=\; \lambda_1\, r_t^{\text{conf}} \;+\; \lambda_2\, r_t^{\text{sum}} \;+\; \lambda_3\, r_t^{\text{stop}} \;+\; r_t^{\text{format}}, \tag{3.15}$$

and the episode return is $R(H) = \sum_{t=1}^{\tau} \gamma^{t-1} r_t$. EAGER requires only answer labels (for $y^\star$) and model scores; it does *not* use frame-level annotations.

Policy Optimization. We optimize π_θ with GRPO [49]. Each iteration, starting from F_1 and p_1, we sample G trajectories $\{H^i\}_{i=1}^G$, compute scalar returns $R(H^i) = \sum_t \gamma^{t-1} r_t^i$, and standardize them to obtain a trajectory-level advantage

$$\hat{A}_i = \frac{R(H^i) - \text{mean}(R(H^\cdot))}{\text{std}(R(H^\cdot))}, \tag{3.16}$$

which is shared across all tokens of trajectory i. Let $H^{i,(n)}$ be the n-th token and N_i the token count in H^i. The GRPO objective is

$$J_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{N_i}\sum_{n=1}^{N_i}\min\Big(\rho_{i,n}\,\hat{A}_i,\; \text{clip}(\rho_{i,n}, 1-\epsilon, 1+\epsilon)\,\hat{A}_i\Big) \tag{3.17}$$

$$\rho_{i,n} = \frac{\pi_\theta(H^{i,(n)} \mid H^{i,<n})}{\pi_{\text{old}}(H^{i,(n)} \mid H^{i,<n})}. \tag{3.18}$$

4. Experimental Studies

We integrate REVISE with both proprietary and open-source VLMs and test it against several baselines. Specifically, we evaluate Qwen2-VL-7B [55], Qwen2.5-VL (3B/7B) [3], InternVL2-8B [8], and GPT-4o [44].

Datasets. We report results on three complementary video-QA benchmarks that probe different temporal scales and reasoning demands. *VideoEspresso* [13] is a large-scale, chain-of-thought (CoT) video reasoning corpus built with a core-frame selection pipeline and multimodal CoT evidence. The benchmark organizes evaluation into 14 tasks spanning causal, temporal, spatial, and high-level narrative reasoning, and emphasizes answering from sparse core frames rather than full streams. *NExT-QA* [66] targets causal and temporal action reasoning with both multiple-choice and open-ended QA. Videos average 44 seconds, and questions are stratified into causal (48%), temporal (29%), and descriptive (23%) types. We report accuracy for multiple-choice following the official split. *EgoSchema* [42] is a long-form egocentric benchmark with >5,000 five-choice multiple-choice questions over 3-minute clips.

Settings. We follow a fixed interaction budget for all multi-turn methods. We set `max_frames_per_round` to 3, i.e., each evidence-gathering step can inspect at most three frames ($\max(F_1, \ldots, F_T) = 3$), and limit the `max_rounds`=4 ($T = 4$). For ablation study, we use VideoEspresso [13] with Qwen2.5-VL-7B [3], for efficiency. Evaluation reports both answer accuracy and the total frame budget consumed by each method. Unless otherwise specified, all VLMs are queried with temperature 0.2, a maximum response length of 256 tokens, and top-p sampling with $p = 0.9$. All VLM experiments are conducted with 4×80G A100.

4.1. Plug-and-Play

REVISE improves any model's video understanding abilities in a "plug-and-play" fashion, without any need to fine-tune the model weights. We compare REVISE with both the proprietary and open-source models across various multiple-choice VQA datasets with several baselines. Our results show that REVISE can achieve comparable performance with far fewer frames.

Improvements with Baselines on VideoEspresso Dataset. As in Table 1, REVISE consistently improves the backbones across both open- and closed-source settings while using only a single-digit number of frames on average. With open-source models, Qwen2-VL [55] + REVISE raises the average from 28.5 to 37.8 (+9.3) and InternVL2 [8] + REVISE from 28.7 to 32.1 (+3.4). On the closed-source side, GPT-4o [44] + REVISE improves the average from 26.4 to 48.9 (+22.5) and achieves the best score in 13 of 14 fine-grained categories, outperforming prior systems such as VideoEspresso. Remarkably, these improvements require only 2.87, 6.25, and 7.99 frames per video for InternVL2, Qwen2-VL, and GPT-4o, respectively, underscoring REVISE's efficiency in sparse, question-focused reasoning.

Table 1. **Comparison of VLMs across fine-grained video reasoning categories.** We report accuracy (%) per category of baseline results, quoted from Han et al. [13] and REVISE under the same protocol of baselines; #Frames reports either the average frames processed per video (ours) or a uniform sampling rate when shown as FPS=k; *Param* is the backbone size. Underlines mark the better score within each backbone pair; bold indicates the group best per column.

Model	#Frames	Param	Narra.	Event	Ingre.	Causal	Theme	Conte.	Influ.	Role	Inter.	Behav.	Emoti.	Cook.	Traff.	Situa.	Avg.
Open-source VLMs																	
LLaVA-1.5 [35]	4	7B	32.3	21.3	19.4	17.1	26.2	20.2	36.1	33.3	21.0	21.1	20.0	35.8	16.7	18.0	24.2
LLaVA-N-Inter [28]	FPS=1	7B	24.2	23.6	26.5	19.2	31.1	32.1	31.9	17.5	24.2	21.1	26.2	30.2	13.3	20.0	24.4
LongVA-DPO [79]	128	7B	35.5	14.9	16.3	19.0	34.4	22.0	37.5	23.8	29.0	22.8	20.0	37.7	16.7	12.0	24.4
mPLUG-Owl3 [70]	FPS=1	7B	30.6	23.6	20.4	22.3	37.7	29.4	48.6	34.9	**30.6**	24.6	27.7	24.5	13.3	24.0	28.0
LLaVA-N-Video [28]	FPS=1	7B	31.2	20.2	16.2	17.6	36.5	32.7	30.6	24.5	26.4	24.5	34.7	20.8	20.3	17.0	25.2
VideoEspresso [13]	2.36	8.5B	**45.2**	27.0	33.7	26.1	39.3	36.7	**55.6**	41.3	**30.6**	**29.8**	30.8	35.8	20.0	26.0	34.1
InternVL2 [8]	FPS=1	8B	33.9	24.1	27.6	24.4	42.6	33.0	45.8	28.6	19.4	22.8	21.5	34.0	20.0	24.0	28.7
InternVL2 + ReViSe	2.87	8B	35.5	29.6	36.0	21.3	39.3	33.0	36.1	31.7	18.3	24.6	36.9	**40.9**	**42.9**	23.4	32.1
Qwen2-VL [55]	FPS=1	7B	27.4	23.0	24.5	23.5	29.5	31.2	47.2	31.7	22.6	28.1	40.0	22.6	30.0	18.0	28.5
Qwen2-VL + ReViSe	6.25	7B	39.2	**39.5**	**38.1**	**33.3**	**50.8**	**46.7**	40.3	**50.0**	25.8	28.1	**43.1**	33.3	36.7	**38.8**	**37.8**
Closed-source VLMs																	
Qwen-VL-Max [2]	FPS=3	–	33.9	22.4	23.5	21.4	26.2	30.3	41.7	30.2	27.4	26.3	20.0	20.8	16.7	24.0	26.0
GPT-4o [44]	FPS=3	–	32.3	16.7	25.5	22.8	32.8	27.5	37.5	28.6	24.2	19.3	30.8	30.2	20.0	22.0	26.4
GPT-4o + ReViSe	7.99	–	**51.9**	**46.5**	**55.1**	**44.0**	**54.1**	**53.2**	**48.6**	**50.8**	**40.3**	**49.1**	**50.8**	**58.5**	**53.3**	**58.0**	**48.9**

Table 2. **Comparison of Training-free or Plug-and-play Methods on EgoSchema (Subset).** Accuracy is reported in %, and efficiency is measured as the number of frames or captions used per video.

Method	Acc. (%)	Frames/Captions Used
VideoAgent [59]	60.2	8.4
VideoTree [63]	66.2	62.4
LVNet [45]	68.2	12
LLoVi [77]	57.6	180
MC-ViT-L [4]	62.6	128+
GPT-4o [44] + REVISE	60.6	9.8

Table 3. **Comparison of training-free or plug-and-play methods on NExT-QA.** Results are quoted from Park et al. [45] for fair comparison. Accuracy is reported as average accuracy (%), and efficiency is measured as the number of frames used per video.

Method	Acc. (%)	Frames/Captions Used
VideoTree [63]	73.5	56
VideoAgent [59]	71.3	8.2
LLoVi [77]	67.7	90
ProViQ [9]	64.6	60
SeViLA [72]	63.6	32
LVNet [45]	61.1	12
GPT-4o [44] + REVISE	63.8	8.4

Comparison across Baselines. We report a comparison with EgoSchema [42] and NExT-QA [66] in Table 2 and Table 3. On EgoSchema (subset), GPT-4o+REVISE achieves 60.6% with 9.8 frames, operating in essentially the same small-budget regime as VideoAgent [59] while avoiding any captioner. Relative to selection- or caption-heavy pipelines, REVISE uses far fewer visual inputs, for example, about 6× fewer than VideoTree [63] and over 13–18× fewer than LLoVi [37] and MC-ViT-L [4], highlighting our emphasis on sparse, direct frame reasoning. On NExT-QA, REVISE achieves 63.8% with 8.4 frames, outperforming LVNet [45] and matching SeViLA [72], while using 3–4× fewer inputs. Compared with the strongest baselines, REVISE trades some accuracy for simplicity and efficiency: it uses ∼6–7× fewer frames than VideoTree and remains in the same ultra-low-frame regime as VideoAgent without relying on caption models. We also evaluate REVISE on additional benchmarks with the same backbone models (Table 12).

Ablation on Frames and Turns. We provide detailed ablation tables in the supplementary. Table 7 shows that increasing the allowed turns consistently lifts accuracy while keeping the frame budget low: the best single-turn setting reaches 38.3% with 4.60 frames, two turns reach 39.0% with fewer frames (3.20), and three turns reach 41.6% at ∼4.0 frames. Allowing four turns yields the best point, 42.1% at just 2.89 frames, while average total rounds remain well below the allowed maximum (≈2.3), indicating early stopping. The accuracy–frames Pareto frontier (Figure 4) is monotonic: as the controller is permitted more turns, it attains strictly better accuracy at strictly lower frame budgets. These trends support our design that multi-round, summary-conditioned selection concentrates evidence into a few targeted frames, improving answer quality without incurring large token costs.

Component Ablation. Table 8 highlights the complementary roles of the persistent summary-as-state and its structured fields ($P/O/H/U/R$). Removing state carryover causes large regressions in accuracy (-18.34%) and nearly doubles computational cost, as the model must repeatedly reconstruct context. Similarly, removing the structured belief/uncertainty/rationale fields leads to substantial drops and the largest runtime increase (+32.14s), indicating that explicit state propagation is critical for stable multi-round reasoning. Ablating both yields the worst accuracy (20.24%). In contrast, the full model achieves the best accuracy while requiring the fewest turns and lowest latency.

Table 4. **Comparison across baseline categories and REVISE (RFT).**

VideoEspresso				
Method	**Acc. (%)**	**Frames**	**Rounds**	**Time (s)**
Direct Reasoning	12.6	8.0	1.00	1.02
Plug-and-Play	20.1	5.2	1.86	1.73
Supervised Format Fine-Tuning	21.3	5.0	1.52	1.67
Reinforced Fine-Tuning	**27.8**	**4.1**	**1.37**	**1.02**
NExT-QA				
Method	**Acc. (%)**	**Frames**	**Rounds**	**Time (s)**
Direct Reasoning	23.6	8.0	1.00	0.88
Plug-and-Play	31.7	5.3	1.74	1.22
Supervised Format Fine-Tuning	27.3	5.1	1.65	1.13
Reinforced Fine-Tuning	**51.3**	**3.9**	**1.32**	**0.62**

4.2. Reinforcement Fine-Tuning

According to Tab. 4, REVISE further enhances the video understanding capabilities of vision-language models (VLMs) when combined with reinforcement fine-tuning. In this setting, the underlying VLM is trained to make better multi-round decisions: what frames to request, how to update its summary-as-state, and when to stop, while keeping the visual encoder fixed.

Training Setup. We employ the multi-round formulation in §3.1 and optimize the policy via GRPO [12, 49]. Specifically, we follow the same setup as RAGEN [62]. We apply our verifier-guided reward (EAGER) that assigns credit based on three signals: (i) confidence gain after new frames are added, (ii) summary sufficiency, measured by whether the final answer is recoverable from the last `<summary>` alone, and (iii) correct-and-early stopping that rewards answering correctly within a small turn budget. The policy is trained on the training split of each dataset using sampled trajectories (1–3 turns, depending on budget), while the summary-as-state is fully externalized and updated at every step. We use Qwen2.5-VL-3B [3] as the backbone model due to computational cost. Because 3B-scale VLMs have limited instruction-following capability, we first distill 8,000 multi-round conversations from GPT-4o [44] under the "plug-and-play" setting and perform a short format fine-tuning stage to teach the model to reliably follow the structured protocol (`<summary>`, `<frames>`, `<answer>`), i.e., output `<summary>` plus either `<frames>` (request) or `<answer>` (final). During reinforcement learning, we adopt the same configuration as Section 4.1, using `max_frames_per_round`=3 and `max_rounds`=4.

Baselines. We compare against three categories of baselines: (i) direct video-understanding models that process all given frames in a single forward pass; (ii) the plug-and-play setting without parameter updates; and (iii) supervised format fine-tuning. All methods are evaluated under a controlled frame budget (typically 3–8 frames per video) to highlight improvements in selection quality and reasoning efficiency.

Metrics. We report answer accuracy, the average number of frames used per example, the number of reasoning rounds, and end-to-end inference time. All methods follow the same output constraints for fair comparison.

Results. Across both datasets, REVISE with reinforcement fine-tuning delivers the strongest performance while simultaneously improving efficiency. On VideoEspresso, REVISE achieves 27.8% accuracy while using only 4.1 frames and 1.37 rounds, outperforming both plug-and-play inference (20.1%) and supervised fine-tuning (21.3%) under the same frame constraints. Notably, REVISE matches the runtime of the single-pass baseline (1.02 s) despite requiring multi-round interaction, demonstrating effective early stopping and tight summary-guided control. On NExT-QA, the gains are even more pronounced: REVISE reaches 51.3% accuracy, surpassing plug-and-play (31.7%) by nearly 20 percentage points, while reducing frames from 5.3 to 3.9 and rounds from 1.74 to 1.32. Reinforcement fine-tuning also cuts inference time nearly in half (0.62 s vs. 1.22 s). These results highlight that REVISE learns to select highly discriminative frames and terminate earlier, achieving substantially better video reasoning ability than both the direct VLM and supervised fine-tuning baselines.

5. Conclusion

We address two challenges in long-video VLM QA: information overload and weak awareness of key evidence. REVISE is a frames-only, multi-round framework that requests a few query-relevant frames, maintains an explicit summary-as-state in `<summary>` with fixed fields $P/O/H/U/R$, and stops early when confident. This design concentrates evidence, keeps token usage low, and preserves cross-turn coherence without modifying the backbone VLM. Across standard VQA benchmarks, REVISE achieves competitive accuracy with single-digit frames on average, often with fewer rounds and lower prompt cost than caption-heavy or dense-frame baselines. Reinforcement fine-tuning further improves open-source VLMs by aligning frame selection, summary quality, and stopping behavior. It yields higher accuracy under the same frame budget. **Limitations.** REVISE still has several limitations. (1) Backbone dependence: performance depends on the underlying VLM's visual fidelity and temporal reasoning, and weaker backbones can degrade the summary state; (2) Interaction latency: multi-round querying introduces additional API calls; single-shot models can be faster under strict latency constraints; (3) Frame granularity: the current system selects whole frames rather than spatial regions, which may limit efficiency on tasks requiring fine-grained localization. Future work could explore adaptive spatial cropping, stronger vision encoders, and extending REVISE to open-ended generation tasks beyond multiple-choice QA.

6. Acknowledgement

CX would like to thank Zhenyu Pan, Jerry Yao-Chieh Hu, Jianshu Zhang, and Lining Mao for the valuable conversation, and Jiayi Wang for facilitating experimental deployments. ZY would like to thank Yanxun Xu for support. The authors would like to thank the anonymous reviewers and program chairs for constructive comments.

HL is partially supported by NIH R01LM1372201, NSF AST-2421845, Simons Foundation MPS-AI-00010513, AbbVie, Dolby, and Chan Zuckerberg Biohub Chicago Spoke Award. This research was supported in part through the computational resources and staff contributions provided for the Quest high-performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 2, 7

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 1, 2, 6, 8

[4] Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. Memory consolidation enables long-context video understanding. *arXiv preprint arXiv:2402.05861*, 2024. 7

[5] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1

[6] Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29071–29082, 2025. 2

[7] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023. 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 6, 7

[9] Rohan Choudhury, Koichiro Niinuma, Kris M Kitani, and László A Jeni. Video question answering with procedural programs. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024. 7

[10] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024. 2

[11] Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. Framemind: Frame-interleaved video reasoning via reinforcement learning. *arXiv preprint arXiv:2509.24008*, 2025. 2

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 3, 8

[13] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191, 2025. 6, 7, 3

[14] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 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*, pages 13504–13514, 2024. 2

[15] Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*, 2025. 2

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[17] Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13702–13712, 2025. 2

[18] De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. Frag: Frame selection augmented generation for long video and long document understanding. *arXiv preprint arXiv:2504.17447*, 2025. 2

[19] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 2

[20] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5627–5646, 2025. 1, 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[22] Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316, 2023. 2

[23] Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *European Conference on Computer Vision*, pages 271–288. Springer, 2024. 2

[24] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024. 3

[25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024. 3

[26] Hosu Lee, Junho Kim, Hyunjun Kim, and Yong Man Ro. Refocus: Reinforcement-guided frame optimization for contextual understanding. *arXiv preprint arXiv:2506.01274*, 2025. 2

[27] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

[28] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 7

[29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2

[30] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

[31] Yunxin Li, Xinyu Chen, Baotain Hu, and Min Zhang. Llms meet long video: Advancing long video question answering with an interactive visual adapter in llms. *arXiv preprint arXiv:2402.13546*, 2024. 2

[32] Yansi Li, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Qiuzhi Liu, Rui Wang, Zhuosheng Zhang, Zhaopeng Tu, Haitao Mi, et al. Dancing with critiques: Enhancing llm reasoning with stepwise natural language self-critique. *arXiv preprint arXiv:2503.17363*, 2025. 3

[33] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984, 2024. 2

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 3

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 7

[36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 1, 2

[37] Licheng Liu, Zihan Wang, Linjie Li, Chenwei Xu, Yiping Lu, Han Liu, Avirup Sil, and Manling Li. A simple" try again" can elicit multi-turn llm reasoning. *arXiv preprint arXiv:2507.14295*, 2025. 3, 7

[38] Haozheng Luo, Ruiyang Qin, Chenwei Xu, Guo Ye, and Zening Luo. Open-ended multi-modal relational reasoning for video question answering. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 363–369. IEEE, 2023. 2

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

[40] Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. S2r: Teaching llms to self-verify and self-correct via reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22632–22654, 2025. 3

[41] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 2

[42] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 1, 6, 7, 3

[43] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 2

[44] OpenAI. GPT-4o blog, 2024. 6, 7, 8

[45] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. In *Proceedings of the 19th Conference of*

the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3569–3588, 2026. 7

[46] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, et al. Detgpt: Detect what you need via reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14172–14189, 2023. 3

[47] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. 2

[48] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[49] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 6, 8, 2

[50] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 1

[51] Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24090–24101, 2025. 2

[52] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898, 2023. 2

[53] Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. Tspo: Temporal sampling policy optimization for long-form video language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9368–9376, 2026. 2

[54] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128, 2025. 2

[55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 6, 7

[56] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In *European Conference on Computer Vision*, pages 142–160. Springer, 2024. 2

[57] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024. 2

[58] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025. 1

[59] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 1, 2, 7

[60] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2023. 2

[61] Yuxuan Wang, Yueqian Wang, Pengfei Wu, Jianxin Liang, Dongyan Zhao, and Zilong Zheng. Lstp: Language-guided spatial-temporal prompt learning for long-form video-text understanding. *arXiv preprint arXiv:2402.16050*, 2024. 2

[62] Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025. 2, 3, 5, 8

[63] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025. 1, 2, 3, 7

[64] Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*, 2025. 3

[65] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024. 2

[66] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 6, 7, 3

[67] 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. 3

[68] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 3

[69] Yifeng Yao, Yike Yun, Jing Wang, Huishuai Zhang, Dongyan Zhao, Ke Tian, Zhihao Wang, Minghui Qiu, and Tao Wang. K-frames: Scene-driven any-k keyframe selection for long video understanding. *arXiv preprint arXiv:2510.13891*, 2025. 2

[70] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 7

[71] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Rethinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8579–8591, 2025. 1, 2

[72] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023. 7

[73] Shoubin Yu, Jaehong Yoon, and Mohit Bansal. Crema: Multimodal compositional video reasoning via efficient modular adaptation and fusion. *arXiv preprint arXiv:2402.05889*, 1, 2024. 2

[74] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning to query frames for video large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[75] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 3

[76] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

[77] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737, 2024. 2, 7

[78] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553, 2023. 2

[79] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 7

[80] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 3

[81] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22056–22065, 2025. 2

[82] Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. *arXiv preprint arXiv:2510.27280*, 2025. 2

[83] Junbo Zou, Ziheng Huang, Shengjie Zhang, Liwen Zhang, and Weining Shen. Videobrain: Learning adaptive frame sampling for long video understanding. *arXiv preprint arXiv:2602.04094*, 2026. 3

[84] Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. Air: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. *arXiv preprint arXiv:2510.04428*, 2025. 3