FORESEEK: SIZING AND FRONT-LOADING MULTIMODAL EVIDENCE FOR LONG-FORM VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form video understanding is a core challenge in multimodal intelligence, where decisive evidence often rests on a few seconds within an hour-long video. Current retrieval-based pipelines typically caption the entire video in advance using an auxiliary vision-language model, while agentic inspection systems gather observations through iterative coarse-to-fine reasoning. Although effective, indexed pipelines incur heavy duration-dependent captioning costs, while iterative agents defer evidence to later turns, by which point the model has already committed to an answer it seldom revises. To overcome these limitations, we introduce ForeSeek, a retrieval-based framework that adaptively sizes and front-loads multimodal evidence before answer commitment. ForeSeek first deploys the Channel-Aligned Evidence Assembler to align grounding frames, sparse visual descriptions, and verbatim speech on a shared timeline via a lightweight retriever. Conditioned on this timeline, the Pre-Answer Evidence Allocator reads model confidence using a fast few-token probe, front-loading a tailored package on the initial pass. Consequently, evidence volume adapts to query demand rather than video duration, confining subsequent interaction to bounded local inspection. Extensive experiments on multiple benchmarks demonstrate the effectiveness of our method, attributable to front-loading adaptively sized evidence packages through pre-answer confidence probing and channel-aligned multimodal assembly.

1 INTRODUCTION

Long-form video understanding is a cornerstone of multimodal intelligence, powering autonomous driving, robotics, and multimodal assistants (Qian et al., 2025; Tian et al., 2025; Zitkovich et al., 2023). Its central challenge is extreme temporal sparsity: within an hour of video holding tens of thousands of frames, decisive clues often last mere seconds (Wang et al., 2025a; Fu et al., 2025; Wu et al., 2024). While multimodal large language models (Bai et al., 2023; Wang et al., 2024a; Liu et al., 2023; Li et al., 2025a) expand context windows, uniform sampling misses transient visual details, whereas dense sampling exceeds context limits. Therefore, providing comprehensive temporal evidence without overwhelming the model remains a fundamental challenge.

To navigate the context limits, existing methods divide into two paradigms depending on evidence timing. Dense indexed pipelines pre-caption entire videos before queries arrive, resolving questions via retrieval over this precomputed index (Ma et al., 2025; Pang & Wang, 2025; Luo et al., 2025; Wang et al., 2025b; Zhang et al., 2025; Yin et al., 2026). While structured, this incurs a heavy video-level captioning cost that scales with duration rather than question demand. Conversely, agentic systems operate on raw video without prior captioning, directing perception through iterative tool calls once queries arrive (Lin et al., 2026; Wang et al., 2024b). Although avoiding upfront costs, they entrust temporal grounding and stopping decisions to coarse-to-fine reasoning loops.

Despite their efficacy, existing paradigms encounter two fundamental constraints. Firstly, dense pipelines answer mainly from visual descriptions, without placing the underlying frames or the speech alongside them on a shared timeline. The model therefore cannot verify one channel against another, even though frames retain details that descriptions omit and speech is cheap to transcribe. Secondly, both paradigms misalign evidence delivery with the model's reasoning dynamics: indexed pipelines pay for full descriptions upfront, whereas agentic systems defer evidence across iterative turns. Crucially, we observe in our own trajectories that the answer committed on the first turn is rarely revised. Therefore, the core challenge is not refining where to look across turns, but assembling complementary evidence and determining its sufficient amount before commitment.

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

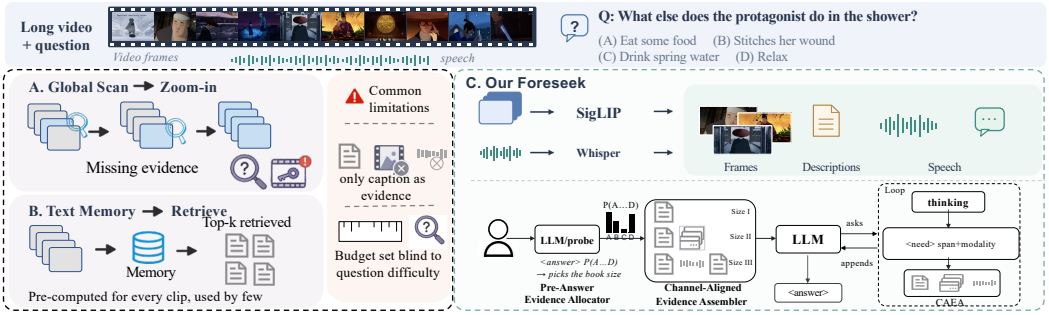

Figure 1: Existing paradigms fix evidence budgets blindly or defer gathering across iterative turns. In contrast, ForeSeek aligns visual frames, descriptions, and verbatim speech on a shared timeline using lightweight encoders, front-loading a tailored package before initial answer commitment. Conditioned on a minimal seed, a fast probe reads predictive confidence to adaptively allocate this upfront budget, reserving subsequent turns strictly for bounded local inspection.

To resolve this challenge, we observe that determining evidence volume is simply choosing a prefix length along a pre-ranked timeline, and that this depth can be diagnosed directly from the model's predictive distribution on a minimal seed, before the model commits to an answer. Grounded in this insight, we design ForeSeek, a retrieval-based framework that sizes and front-loads multimodal evidence before the model commits to an answer. As illustrated in Fig. 1, ForeSeek realizes this strategy through two core modules: the Channel-Aligned Evidence Assembler and the Pre-Answer Evidence Allocator. First, the Channel-Aligned Evidence Assembler aligns grounding frames, sparse visual descriptions, and verbatim transcripts of the spoken audio on a shared timeline using a lightweight retriever without large language models. By scoring visual and acoustic cues along a unified curve, it constructs prefix-nested packages that scale text volume without shifting visual anchors. Second, the Pre-Answer Evidence Allocator determines the appropriate evidence tier directly from the internal confidence of the answering model. Conditioned on the minimal seed, a fast four-token probe reads the model's likelihood over candidate options, maps it to a compact budget of ten, twenty, or thirty moments, and delivers the full package in the very first pass. By coordinating evidence assembly with predictive allocation, ForeSeek replaces unbounded video-level captioning with an adaptive question-level budget, while restricting subsequent interaction to bounded local inspection.

We extensively validate the effectiveness of ForeSeek on hour-scale benchmarks including LVBench and Video-MME, as well as complex reasoning scenarios on Video-Holmes. Our framework outperforms previous methods across multiple evaluation metrics on these datasets while drastically reducing evidence overhead. Our contributions can be summarized as follows:

- We show that front-loading multimodal evidence before answer commitment avoids high offline captioning costs.
- We introduce ForeSeek, a novel framework that integrates pre-answer confidence probing and timeline-aligned evidence front-loading to eliminate unnecessary deliberation loops.
- We validate the effectiveness of our framework across multiple benchmarks, establishing advanced results with significant evidence savings.

2 Related Work

Long Video Understanding. Long-video benchmarks (Wang et al., 2025a; Fu et al., 2025; Wu et al., 2024; Cheng et al., 2025) evaluate retrieval, temporal grounding and reasoning over clues spread across videos far longer than a single forward pass can cover. Video MLLMs (Li et al., 2025a; Wang et al., 2024a; Bai et al., 2025b) answer in a single pass, and long-video variants either raise how many frames that pass can absorb through token compression and long-context training (Shu et al., 2025; Shen et al., 2025; Chen et al., 2025c) or select the frames most relevant to the question (Tang et al., 2025). Beyond the single pass, two families have emerged. Agentic methods operate on the raw video and gather observations through iterative tool calls (Wang et al., 2024b; Yang et al., 2025; Chen et al., 2025a; Yuan et al., 2026; Wang et al., 2026b; Lin et al., 2026; Zhang et al., 2026; Yin et al., 2026; Wang et al., 2026a): a language model decides at each step which interval to expand and when to stop, with vision–language models serving as tools that retrieve and describe frames. Indexed pipelines instead describe the video before any question is asked and answer by retrieval

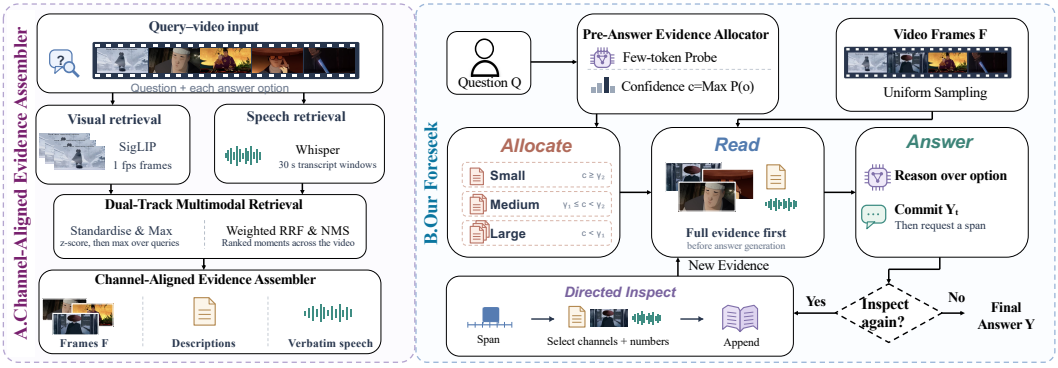

Figure 2: The Channel-Aligned Evidence Assembler scores visual frames and speech transcripts using lightweight encoders, constructing prefix-nested evidence books alongside invariant visual grounding frames. Conditioned on a minimal seed, the Pre-Answer Evidence Allocator reads predictive confidence through a few-token probe to deliver a tailored evidence package before generation begins. The answering model then commits to an initial prediction on the first turn, resolving subtle details through bounded local inspections before producing the final answer.

over the stored descriptions (Ma et al., 2025; Wang et al., 2025b; Pang & Wang, 2025; Luo et al., 2025; Zhang et al., 2025), which makes the description cost scale with video duration rather than with what a question needs. Individual systems are detailed in App. A.6. Unlike these pipelines, ForeSeek ranks moments on visual and spoken content without a large language model, describes only the moments it selects, and pairs them with frames and verbatim transcripts.

Adaptive Evidence Acquisition. Several methods adapt evidence acquisition to the question. Route2Look (Wang et al., 2026a) routes among acquisition tools by query type, and agentic loops stop once the model judges its evidence sufficient (Wang et al., 2024b; Zhang et al., 2025; Lin et al., 2026). In text question answering, FLARE (Jiang et al., 2023) retrieves when upcoming tokens have low probability, Self-RAG (Asai et al., 2024) learns to emit its own retrieval decisions, and Adaptive-RAG (Jeong et al., 2024) routes questions by predicted complexity. Our work differs in reading the model's confidence over the options before it answers and using it to set how much multimodal evidence each question starts with.

3 Methodology

The amount of evidence needed to answer a long-video question varies across queries, motivating a question-specific budget before initial answer commitment. Our key observation is that this budget can be expressed as a prefix length over ranked candidate moments and selected using the model's predictive confidence on a minimal evidence seed. Building on this observation, we present **ForeSeek**, a retrieval-based framework that sizes and front-loads multimodal evidence before the first full reasoning-and-answer pass. As illustrated in Fig. 2, ForeSeek coordinates two core modules. The **Channel-Aligned Evidence Assembler** in Sec. 3.2 aligns grounding frames, visual descriptions and verbatim transcripts on a shared timeline, constructing nested evidence books over a frame sweep that does not scale with the book. The **Pre-Answer Evidence Allocator** in Sec. 3.3 uses a few-token probe on the minimal seed to select the book size for each question. The model then reasons over the selected package, with **Bounded Local Inspection** in Sec. 3.4 providing additional evidence for unresolved details. The answering model is used off the shelf without parameter updates.

3.1 Preliminaries

Problem Formulation. Given an hour-scale video $\mathbf{X}$, a question $\mathbf{Q}$, and options $O = \{o_1, \ldots, o_K\}$, the goal of long-video question answering is to predict the correct answer $\mathbf{Y} \in O$ using an off-the-shelf model p_θ. Existing agentic systems generate a multi-turn trajectory τ of thought, action, and observation steps, deciding where to look and how much evidence to gather within the same loop. In contrast, ForeSeek separates evidence volume determination from subsequent span inspection. Before generating an answer, the model receives an evidence package:

$$\mathbf{E}_M = \langle \mathbf{F}, \mathcal{B}_M \rangle, \tag{1}$$

where $\mathbf{F}$ denotes invariant visual grounding frames, and $\mathcal{B}_M$ combines sparse visual descriptions and verbatim transcripts across M retrieved moments. ForeSeek assigns each question an initial

evidence tier $M^\star \in \{M_{\text{small}}, M_{\text{mid}}, M_{\text{large}}\}$ through a probe before initial answer commitment, balancing coverage against reading and captioning costs. Conditioned on $\mathbf{E}_{M^\star}$, the model commits to an initial answer, submits a mandatory span request for missing details, and executes at most H inspection turns returning localized frames, descriptions, or transcripts.

Action Space and Toolset. To balance multimodal evidence with computational cost, ForeSeek strictly separates framework-level evidence assembly from model-level action execution, summarized in Tab. 6. Unlike standard agentic systems that burden the model with iterative video search, ForeSeek gathers global evidence before the model commits to an answer. The framework assembles and sizes this upfront package before the model commits to an answer, reading the model's own confidence with a BUDGETPROBE that decodes only a few tokens and requires no auxiliary scorer. This leaves only two primitives in the model's action space: INSPECT for localized inspection and ANSWER for final prediction, converting open-ended search into bounded inspection. Their operational behaviors are defined as follows:

- **GLOBALBROWSE**: Samples F frames across a uniform temporal grid over the video without the query, providing a holistic visual baseline and covering segments the retriever does not select.
- **SEMANTICRETRIEVE**: Computes cross-modal relevance between queries and moments across visual and spoken channels, outputting temporally dispersed top candidates.
- **BUDGETPROBE**: Conditions the off-the-shelf model on the minimal seed, decodes a few tokens to assess candidate option likelihoods, and maps this confidence to an assigned budget tier $M^\star$.
- **INSPECT**: Takes a temporal span, a channel mask, and a frame count, returning the frames, clip descriptions, or verbatim spoken lines within that window, up to a maximum span.
- **ANSWER**: Emits the final option prediction and terminates the reasoning trajectory.

3.2 CHANNEL-ALIGNED EVIDENCE ASSEMBLER

Visual frames, clip descriptions, and spoken transcripts provide complementary evidence for long-video questions. The Channel-Aligned Evidence Assembler brings these channels onto a shared timeline, using lightweight dual-track retrieval to select moments for textual packaging while keeping the grounding frames fixed across evidence tiers.

Dual-Track Multimodal Retrieval. Given the question $\mathbf{Q}$ and candidate options $O = \{o_1, \ldots, o_K\}$, we form an expanded query set $\mathcal{Q} = \{\mathbf{Q}\} \cup \{o_k\}_{k=1}^{K} \cup \{\mathbf{Q} \circ o_1 \circ \cdots \circ o_K\}$, where $\circ$ denotes textual concatenation. This formulation ensures that a video moment receives a high score if it correlates with the question alone, with any specific answer hypothesis, or with the question and its full option set taken together. The video is processed along two parallel tracks: a visual track operating on decoded frames x_t sampled at one frame per second, and a speech track operating on transcript segments w_t extracted across thirty-second windows via Whisper. We employ an image-text encoder ϕ, specifically SigLIP, for the visual track and a sentence encoder ψ, specifically BGE, for the speech track. For each second t, the relevance scores on the visual and speech tracks are computed as:

$$\tilde{s}_t^v = \max_{q \in \mathcal{Q}} z\big(\cos(\phi(x_t), \phi(q))\big), \qquad \tilde{s}_t^a = \max_{q \in \mathcal{Q}} z\big(\cos(\psi(w_t), \psi(q))\big), \tag{2}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity, and $z(\cdot)$ denotes intra-video standardization applied per query, which transforms the score distribution of each query to zero mean and unit variance. Standardizing within a track places the different queries on a comparable scale, so that the maximum over $\mathcal{Q}$ reflects genuine relevance rather than a query whose cosine values are intrinsically higher. The two tracks are then fused by weighted reciprocal-rank fusion, which compares them by position rather than by magnitude:

$$\bar{s}_t = \sum_{c \in \{v,a\}} \frac{w_c}{\kappa + r_c(t)}, \tag{3}$$

where $r_c(t)$ is the rank of second t within track c counted from zero, w_c weights the contribution of each track, and κ damps the influence of the top ranks. Ranking makes the fusion invariant to the scale of either similarity, and the speech track is given the smaller weight because speech covers only part of the video. To prevent candidate moments from clustering around a single salient event, we apply non-maximum suppression along the timeline, which yields a temporally dispersed sequence of ranked candidate moments $\mathcal{P} = (t_1, t_2, \ldots)$.

Prefix-Nested Evidence Packaging. For an allocated evidence budget M, the assembler gathers the top-M candidate moments from $\mathcal{P}$ to construct the multimodal evidence book $\mathcal{B}_M$. Each moment contributes two aligned textual channels: the description of the clip that contains it, and the verbatim speech transcript spoken during that interval. These entries are ordered chronologically, presenting

the model with an aligned multimodal transcript of salient events rather than a ranked list. Crucially, because every tier is read off the same ranking $\mathcal{P}$ under the same suppression window, the books are strictly nested:

$$\mathcal{B}_{M_{\text{small}}} \subset \mathcal{B}_{M_{\text{mid}}} \subset \mathcal{B}_{M_{\text{large}}}. \tag{4}$$

Expanding the evidence tier therefore inserts further moments at their chronological positions along the timeline without altering or displacing any evidence already delivered.

Tier-Invariant Visual Anchors. The visual grounding frame set $\mathbf{F}$ is decoupled from the textual tier M. It comprises F frames sampled on a uniform temporal grid across the entire video, chosen without reference to the question. The uniform grid guarantees that no stretch of the video is absent from what the model sees first, including stretches the retriever left unscored; frames at retrieved moments are not added to it, since the model can request them through INSPECT. Crucially, $\mathbf{F}$ remains identical across all evidence tiers $\mathcal{B}_{M_{\text{small}}}$, $\mathcal{B}_{M_{\text{mid}}}$ and $\mathcal{B}_{M_{\text{large}}}$. Consequently, the probe of Sec. 3.3, computed on $\langle \mathbf{F}, \mathcal{B}_{M_{\text{small}}} \rangle$, reads the model's confidence on the very frames it will answer with, and switching evidence tiers alters only the depth of written evidence, so the effect of the budget can be measured on its own. With the evidence books defined, the remaining question is how deeply each query needs to draw from the ranked candidates.

3.3 PRE-ANSWER EVIDENCE ALLOCATOR

To determine this depth, the Pre-Answer Evidence Allocator reads the model's predictive confidence on the smallest evidence book and maps it to one of three tiers. The choice is made before the first answer is written because a model seldom revises an answer once committed (Sec. 3.4).

Confidence Probe on the Smallest Book. The allocator constructs the minimal input $\mathbf{E}_{\text{small}} = \langle \mathbf{F}, \mathcal{B}_{M_{\text{small}}} \rangle$ with the question $\mathbf{Q}$ and its K options $\mathcal{O}$, prefills the opening answer marker, and decodes only a few tokens. At the decoded position where the option letters carry the most probability mass, the probability of each option $o \in \mathcal{O}$ is read and re-normalized over the option set, giving $p(o)$. The pre-answer confidence is the maximum option likelihood,

$$c = \max_{o \in \mathcal{O}} p(o), \tag{5}$$

obtained before the model commits to an answer, with $c \in [1/K, 1]$ by construction. Decoding a few tokens keeps the probe's overhead negligible; full reasoning and answering follow in a separate pass after allocation.

Predictive Three-Tier Allocation. To map the confidence c to an evidence budget, two thresholds $\gamma_1 < \gamma_2$ partition its range into three tiers:

$$M^\star = M_{\text{large}} \text{ if } c < \gamma_1, \quad M_{\text{mid}} \text{ if } \gamma_1 \le c < \gamma_2, \quad M_{\text{small}} \text{ if } c \ge \gamma_2, \tag{6}$$

where $M_{\text{small}} < M_{\text{mid}} < M_{\text{large}}$ are the three book sizes. High confidence keeps the smallest book and its short prompt; low confidence triggers the largest to maximize recall; intermediate scores receive M_{mid}. Because the books are nested (Eq. 4), moving to a larger tier only adds written evidence to $\mathcal{B}_{M^\star}$ while the frames $\mathbf{F}$ stay invariant.

Upfront Evidence Delivery. Once the budget tier $M^\star$ is determined, the framework delivers the tailored package $\mathbf{E}_{M^\star} = \langle \mathbf{F}, \mathcal{B}_{M^\star} \rangle$ in full before the answering model writes a single reasoning token. Front-loading the complete package lets the first answer rest on all the allocated evidence, reserving later interaction for Bounded Local Inspection in Sec. 3.4.

3.4 BOUNDED LOCAL INSPECTION WORKFLOW

Once the Pre-Answer Evidence Allocator delivers $\mathbf{E}_{M^\star}$, downstream reasoning proceeds through **Bounded Local Inspection**. The upfront package provides sufficient global context for an initial commitment, but not every question is settled by it. Unlike agentic frameworks that run an open-ended search over the video, ForeSeek confines downstream reasoning to inspecting localized spans, requested only after the model has committed to an answer and within a bounded budget.

First-Turn Committed Reasoning. In the first turn, the answering model receives the initial trajectory $\tau_1 = \langle \mathbf{F}, \mathcal{B}_{M^\star}, \mathbf{Q} \rangle$. The model evaluates candidate options against this evidence, eliminates refuted choices, and commits to an initial prediction $\mathbf{Y}_1$. In the same turn, it issues a targeted span request a_1 to inspect the details that separate the remaining candidates. To prevent premature termination, this initial inspection is mandatory; if the model omits a span, the framework prompts it once with directive I_{req}, asking for the span whose evidence would most change its answer were it wrong. Because the answer is written before any request, every later inspection is measured against a stated commitment rather than an open hypothesis.

Bounded Iterative Verification. Given a span request a_t, the framework executes INSPECT$(\mathbf{X}, a_t)$ to fetch localized frames, clip descriptions, or the verbatim transcript v_t, which is appended to the

Table 1: Main comparison results on three long video understanding benchmarks (Acc., %, ↑ and Frames, ↓). Both ForeSeek rows adaptively allocate discrete evidence packages via confidence probing. Best results are in **bold**, and second-best are underlined.

Method	Reference	LVBench		VideoMME (Long)		LongVideoBench (Long)	
		Frames ↓	Acc ↑	Frames ↓	Acc ↑	Frames ↓	Acc ↑
VLM-based methods							
Qwen2.5-VL-72B (Bai et al., 2025b)	[ArXiv2025]	768	47.3	256	53.2	-	-
Qwen3-VL-8B (Bai et al., 2025a)	[ArXiv2025]	768	45.8	256	59.3	-	-
Qwen3.5-9B (Qwen Team, 2026a)	[Blog2026]	768	60.8	256	67.3	-	-
Gemini-1.5-Pro (Gemini Team et al., 2024)	[ArXiv2024]	3600	33.1	1233	67.4	256	58.6
Gemini 2.0 Flash (Pichai et al., 2024)	[Blog2024]	4037	48.3	1233	63.0	256	45.7
GPT-4o (Hurst et al., 2024)	[ArXiv2024]	384	30.8	384	65.3	256	60.9
GPT-5 (Singh et al., 2025)	[ArXiv2025]	384	60.1	384	67.9	384	64.5
Qwen3.8-27B (Qwen Team, 2026b)	[Blog2026]	384	62.2	384	69.2	384	68.8
GLM-5.3-Flash (Z.ai, 2026)	[Blog2026]	384	65.9	384	71.8	384	64.2
Agent-based methods							
VideoAgent (Wang et al., 2024b)	[ECCV2024]	25.5	29.3	24.6	46.4	-	-
VideoTree (Wang et al., 2025b)	[CVPR2025]	103.2	28.8	98.0	53.1	-	-
DrVideo (Ma et al., 2025)	[CVPR2025]	-	-	493.2	51.7	-	-
VCA (Yang et al., 2025)	[ICCV2025]	20.0	41.3	18.1	54.2	-	-
Mr.Video (Pang & Wang, 2025)	[NeurIPS2025]	8074	60.8	4932	61.8	2816	61.6
DVD (Zhang et al., 2025)	[NeurIPS2025]	8074	74.2	4932	67.3	2816	68.6
VideoSeek (Lin et al., 2026)	[CVPR2026]	92.3	68.4	60.9	70.1	29.6	73.5
Route2Look (Wang et al., 2026a)	[EMNLP2026]	202.3	75.4	126.6	76.1	168.7	**77.8**
Ours							
ForeSeek (Qwen3.8-27B)		117	75.7	107	80.6	75	75.5
ForeSeek (GLM-5.3-Flash)		110	**78.5**	107	**83.1**	109	76.8

trajectory τ. In subsequent turns, the model re-evaluates its reasoning, updates its answer commitment to $\mathbf{Y}_t$, and either requests another span or terminates. To curb visual computation and redundancy, frame requests are subject to two guards. A span overlapping previously inspected windows in $\mathcal{S}$ by more than half is refused once and the model asked to re-target; a second such request ends the interaction. A request that would raise the cumulative span of fetched frames above a fraction ρ of the video ends it as well. The interaction also terminates when the model requests no further spans or reaches the limit of H inspection steps, returning the last committed prediction $\mathbf{Y}$.

Design Rationale. Bounded Local Inspection balances evidence sufficiency against deliberation cost. The answer committed on the first turn is rarely revised in our trajectories(Sec. 3.4), so evidence supplied after the first answer rarely changes it. ForeSeek therefore front-loads $\mathbf{E}_{M^\star}$ so that the first answer rests on all the allocated evidence, and reserves subsequent turns for localized inspection. Selective inspection lets the model choose the window, the modalities and the frame count of each request, consuming fewer frames than a fixed fetch policy at equal accuracy. The guards bound how far the context can grow, so the inference cost of a question is bounded as well.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate ForeSeek on three long-video understanding benchmarks and one complex video reasoning benchmark, all formulated as multiple-choice question answering.

- **LVBench** (Wang et al., 2025a) contains 1,549 questions over 103 videos averaging more than an hour. It covers six tasks: entity recognition, event understanding, key information retrieval, temporal grounding, reasoning, and summarization.
- **Video-MME** (Fu et al., 2025) evaluates video understanding across different durations. We use its long subset, comprising 900 questions over 300 videos lasting 30–60 minutes.
- **LongVideoBench** (Wu et al., 2024) evaluates understanding of extended video contexts. We use its long validation subset, containing 564 questions over 188 videos lasting 15–60 minutes.
- **Video-Holmes** (Cheng et al., 2025) evaluates multi-clue reasoning over extended contexts. It contains 1,837 questions across 270 suspense short films, spanning seven specialized reasoning tasks such as social reasoning, causal inference, and timeline analysis.

Implementation Details. Our framework uses off-the-shelf models without task-specific tuning.

Table 2: Multi-clue reasoning results on Video-Holmes (Acc., %, ↑ and Frames, ↓). † marks frame counts estimated at 1 FPS. Best results are in **bold**, and second-best are underlined.

Method	Reference	Frames ↓	SR	IMC	TCI	TA	MHR	PAR	CTI	Overall ↑
VLM-based methods										
Qwen2.5-VL-32B (Bai et al., 2025b)	[ArXiv2025]	32	43.2	44.2	31.5	51.0	36.4	31.4	32.2	38.4
SEED-Bench-R1 (Chen et al., 2025b)	[ArXiv2025]	32	42.8	35.1	25.6	40.5	29.2	29.9	32.6	33.5
VideoChat-R1 (Li et al., 2025b)	[ArXiv2025]	32	42.1	38.8	24.5	39.5	29.5	27.8	29.3	33.0
Video-R1 (Feng et al., 2025)	[NeurIPS2025]	32	48.6	41.7	28.9	34.5	31.0	33.5	35.9	36.5
GPT-4o (Hurst et al., 2024)	[ArXiv2024]	32	50.0	49.6	38.8	30.0	44.0	39.2	37.0	42.0
Gemini 1.5 Pro† (Gemini Team et al., 2024)	[ArXiv2024]	185.1	52.1	48.2	34.4	26.0	39.2	46.4	38.9	41.2
Gemini 2.5 Pro† (Comanici et al., 2025)	[ArXiv2025]	185.1	46.6	49.3	46.9	53.0	40.1	44.3	37.4	45.0
Gemini 2.0 Flash† (Pichai et al., 2024)	[Blog2024]	185.1	41.8	33.7	23.1	20.5	30.1	26.8	33.7	30.6
Gemini 2.0 Flash Thinking† (Basu Mallick & Kilpatrick, 2025)	[Blog2025]	185.1	43.4	46.9	43.1	51.0	37.9	43.6	39.3	43.1
GPT-5 (Singh et al., 2025)	[ArXiv2025]	384	47.2	43.4	40.6	53.5	46.3	38.1	39.6	44.1
Qwen3.8-27B (96 frames) (Qwen Team, 2026b)	[Blog2026]	96	63.0	54.0	50.9	49.0	56.3	50.0	52.2	54.2
GLM-5.3-Flash (384 frames) (Z.ai, 2026)	[Blog2026]	384	59.9	63.4	55.3	50.5	59.3	55.2	48.9	56.5
Agent-based methods										
VideoSeek (Lin et al., 2026)	[CVPR2026]	42.7	56.1	43.8	45.0	54.5	46.6	43.3	41.8	47.3
Our method										
ForeSeek (Qwen3.8-27B)		94	**68.8**	63.0	**60.8**	61.0	60.8	**60.8**	**53.3**	61.4
ForeSeek (GLM-5.3-Flash)		96	66.7	**70.3**	**60.8**	**68.5**	**64.5**	**60.8**	49.3	**63.0**

- **Answering models.** We deploy Qwen3.8-27B (Qwen Team, 2026b) as the primary backbone and GLM-5.3-Flash (Z.ai, 2026) to evaluate model transfer. Each backbone independently probes its predictive confidence to select the initial evidence tier.
- **Evidence preparation.** We employ SigLIP for visual retrieval, BGE for transcript search, and Whisper to transcribe audio across all benchmarks, discarding external subtitles for consistency.
- **Evidence allocation.** Initial evidence books are selected from moment tiers using a few-token confidence probe with thresholds $\gamma_1 = 0.6$ and $\gamma_2 = 0.9$, keeping grounding frames fixed.

Evaluation Metrics. We report answer accuracy (Acc., %, ↑) across benchmarks and per-task breakdown on Video-Holmes. Visual input is measured by average frames delivered per question (Frames, ↓), covering initial context and subsequent inspections for ForeSeek.

4.2 Hour-Scale Video Understanding on LVBench

As reported in Table 1, we evaluate hour-scale video understanding on LVBench, where videos average more than an hour and require pinpointing dispersed multimodal evidence. Equipped with the Qwen3.8-27B backbone, ForeSeek establishes superior performance, achieving an overall accuracy of 75.7%. This result demonstrates a decisive margin of 13.5 percentage points over direct answering with the identical backbone at 62.2%, while concurrently reducing the visual input from 384 frames to a compact average of 117 frames per question. When compared against specialized video agents, ForeSeek demonstrates substantial advantages across distinct operational paradigms. Against the strongest adaptive agent, Route2Look at 75.4% accuracy with 202.3 frames, ForeSeek achieves higher accuracy while consuming only 58% of its visual frame budget. Furthermore, ForeSeek surpasses the dense indexing system DVD at 74.2% by 1.5 percentage points, completely eliminating the extreme computational burden of processing thousands of frames per video. It also outpaces exploratory tool-use agents such as VideoSeek at 68.4% by a clear margin of 7.3 percentage points. We attribute this dual advantage in accuracy and efficiency to the front-loading of adaptively sized evidence packages, which anchors global reasoning prior to commitment and confines downstream interaction strictly to bounded local inspection.

The framework also transfers effectively across distinct model families without architectural modification. Driven by the GLM-5.3-Flash backbone, ForeSeek attains the top accuracy of 78.5% with only 110 frames per question on the evaluation dataset documented in Tab. 1, exceeding its corresponding direct-answering baseline of 65.9% by 12.6 percentage points. These results confirm that calibrating evidence volume through pre-answer confidence probing provides a reliable and scalable mechanism for hour-scale video understanding.

4.3 Multimodal Long-Video Understanding on Video-MME

As summarized in Table 1, we evaluate multimodal long-video comprehension on the long subset of Video-MME, where videos range from 30 to 60 minutes and require sustained reasoning across vision and speech. ForeSeek with Qwen3.8-27B achieves an advanced accuracy of 80.6%, exceeding its direct-answering backbone baseline at 69.2% by 11.4 percentage points while using only 107 frames per question, approximately 28% of the baseline's 384 frames. Compared with specialized

Table 3: Ablation of core evidence components on LVBench.

Frames	Desc	Trans	Acc	Text	Frames
✓	-	-	63.1	0	130
-	✓	-	69.0	21.8	0
✓	✓	-	72.6	21.4	119
✓	✓	✓	**75.7**	29.0	117

Table 4: Ablation of pre-answer routing on LVBench.

Routing	Acc	Text	Frames
-	74.8	29.3	118
✓	**75.7**	29.0	117

Table 5: Evidence packet size against cost.

Size	Acc	Frames	Tokens
10	72.1	120.0	4.5k
20	74.8	117.5	7.7k
30	74.8	115.9	10.9k
Ours	**75.7**	116.8	7.6k

agents, ForeSeek outperforms Route2Look at 76.1% accuracy with 126.6 frames by 4.5 percentage points under a lighter visual budget, and exceeds VideoSeek at 70.1% and DVD at 67.3% by 10.5 and 13.3 percentage points, respectively, without the heavy burden of processing thousands of frames per video. These gains stem from the Channel-Aligned Evidence Assembler, which synchronizes visual grounding frames with clip descriptions and verbatim speech transcripts along a unified timeline, securing rich cross-modal evidence while circumventing exploratory search drift. Under the GLM-5.3-Flash backbone, ForeSeek attains the leading accuracy of 83.1% with 107 frames per question, surpassing its direct-answering counterpart at 71.8% by 11.3 percentage points. These results corroborate that front-loading adaptively allocated multimodal packages provides a reliable foundation for long-video reasoning across distinct reasoning engines.

4.4 Long-Context Video Question Answering on LongVideoBench

We further examine extended-context understanding on LongVideoBench in Table 1, where videos span 15 to 60 minutes. Under GLM-5.3-Flash, ForeSeek attains 76.8% accuracy with 109 frames, outperforming direct answering at 64.2% by 12.6 percentage points. Compared with specialized agents, it rivals Route2Look at 77.8% while requiring 35% fewer frames, and outpaces VideoSeek and DVD by 3.3 and 8.2 percentage points. With Qwen3.8-27B, ForeSeek reaches 75.5% accuracy, gaining 6.7 percentage points over direct answering while cutting frames by 80% from 384 to 75. It also surpasses VideoSeek and DVD by 2.0 and 6.9 percentage points using only 3% of DVD's frames. These gains confirm that front-loading grounded multimodal evidence resolves extended narrative threads without context dilution or heavy visual overhead.

4.5 Multi-Clue Video Reasoning on Video-Holmes

Video-Holmes evaluates multi-clue synthesis and temporal deduction across suspense films, shifting from coverage to multi-step reasoning. In Table 2, ForeSeek achieves top overall accuracies of 61.4% with Qwen3.8-27B and 63.0% with GLM-5.3-Flash, outperforming direct answering by 7.2 and 6.5 percentage points using only 94 and 96 frames, while surpassing VideoSeek at 47.3% by over 14 percentage points. On tasks requiring dialogue and subtle visual cues, ForeSeek achieves 68.8% on social reasoning and 63.0% on intention chaining with Qwen3.8-27B, while GLM-5.3-Flash reaches 70.3% on intention chaining, outperforming VideoSeek by up to 26.5 percentage points. Both backbones also attain 60.8% on temporal causal inference and physical anomaly reasoning, confirming that channel-aligned evidence front-loading uncovers dispersed narrative clues.

4.6 Ablation Study

To systematically dissect the algorithmic components of ForeSeek, we conduct extensive ablation experiments on LVBench with Qwen3.8-27B as the default answering model.

Effect of core evidence components. To evaluate each multimodal channel in ForeSeek, Tab. 3 systematically ablates grounding frames, visual descriptions, and speech transcripts on LVBench. Frames alone yield the lowest accuracy of 63.1%, as uniform sweeping frequently misses decisive seconds. Replacing frames with written descriptions recovers accuracy to 69.0% without visual input, demonstrating that semantic grounding matters more than raw visual density. Combining descriptions with grounding frames elevates performance to 72.6%, confirming that text and visual details offer complementary cues. Finally, integrating verbatim speech transcripts achieves our full accuracy of 75.7%, supplying dialogue evidence that visual channels omit.

Efficacy of pre-answer routing. Tab. 4 isolates the Pre-Answer Evidence Allocator by comparing performance with and without routing under fixed evidence components. Both settings supply grounding frames, visual descriptions, and speech transcripts. Adaptive routing improves accuracy from 74.8% to 75.7%, while reducing frames from 118 to 117 and text from 29.3k to 29.0k characters. By probing confidence before answering, the allocator tailors evidence to question difficulty, avoiding over-provision on simple questions while ensuring sufficient context for complex queries.



Figure 3: **Case studies across question types.** Six LVBench questions answered by Qwen3.8-27B, one per question type, with every line inside a panel quoted verbatim from the run log. The band at the top of each panel gives the question type, the probe's reading over the option letters, the packet size that reading selects and the number of rounds the question took; the correct option is set in bold.

Evidence budget scaling and cost analysis. Tab. 5 benchmarks fixed small, mid, and large books of 10, 20, and 30 retrieved moments against adaptive allocation with Qwen3.8-27B, all on the same 1,549 LVBench questions with 115.9 to 120.0 frames. Returns on written evidence saturate rapidly: expanding from small to mid gains 2.7 percentage points from 72.1% to 74.8% for 3.2k tokens, whereas expanding to large spends 42% more tokens without further gain. Adaptive allocation reaches 75.7% using 7.6k tokens, 30% fewer than the large tier, and avoids the 3.6 percentage point degradation at the small tier, the only statistically significant difference ($p = 1.4 \times 10^{-4}$). Rather than a static budget, the confidence probe sets the evidence volume per question.

Qualitative analysis across question types. Fig. 3 shows representative reasoning trajectories across LVBench question types. For event understanding and information retrieval, the routed packet settles questions without additional inspection. Temporal grounding and summarization require one inspection round, pairing transcripts with frames for spoken and captioned counts. Reasoning and entity recognition take two rounds to identify entities via transcripts and verify visual details. Competing methods fail on at least one question while ForeSeek succeeds across all six. These cases confirm that evidence demand is an intrinsic question property satisfied by adaptive allocation.

5 Conclusion

We introduce **ForeSeek**, a retrieval-based framework for long-form video understanding that sizes and front-loads multimodal evidence before initial answer commitment. ForeSeek couples two components: the Channel-Aligned Evidence Assembler, which gathers grounding frames, sparse descriptions, and verbatim transcripts without model querying, and the Pre-Answer Evidence Allocator, which sizes evidence tiers via confidence probing. Evaluations across four benchmarks demonstrate that upfront evidence decisively outperforms uniform frame sweeps while preventing search drift. We discuss limitations and future directions in Appendix A.7.

AI USE STATEMENT

In this work, we have not used generative AI tools to implement methods, to generate synthetic data sets, to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, to clean or reformat data sets, to support qualitative or thematic data analysis, or to interpret results; formulating mathematical claims, providing ingredients for or assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, to edit the paper to improve readability, and to create or modify scientific figures. We have reviewed all AI-assisted work: AI-generated code was read, verified and tested by the authors before use, every passage of AI-edited text was checked by the authors against the underlying experiments, and AI-assisted figures were checked against the data they display. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies multimodal long-video understanding and temporal reasoning on public academic benchmarks, including LVBench, Video-MME, LongVideoBench, and Video-Holmes. All datasets and video materials are sourced from publicly available benchmarks and utilized strictly in accordance with their respective research licenses. This research does not involve human subjects, private personal data, or potential dual-use applications beyond standard multimodal artificial intelligence research.

REPRODUCIBILITY STATEMENT

All methodological formulations, retrieval configurations, and experimental protocols are thoroughly documented in the main text and appendix. We will release the complete codebase, prompt templates, and evaluation pipelines upon acceptance.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. 3

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. *arXiv preprint arXiv:2308.12966*, 2023. 1

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. 6, 16

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. 2, 6, 7, 16

Shrestha Basu Mallick and Logan Kilpatrick. Gemini 2.0: Flash, Flash-Lite and Pro. https://developers.googleblog.com/en/gemini-2-family-expands/, 2025. Accessed 2026-09-25. 7

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. LVAgent: Long video understanding by multi-round dynamical collaboration of MLLM agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20237–20246, 2025a. 2

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. Exploring the effect of reinforcement learning on video understanding: Insights from SEED-Bench-R1. *arXiv preprint arXiv:2503.24376*, 2025b. 7

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. LongVILA: Scaling long-context visual language models for

long videos. In *International Conference on Learning Representations (ICLR)*, pp. 18227–18246, 2025c. 2

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM think like Holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. 2, 6, 16

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99114–99137, 2025. 7

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. 1, 2, 6, 16

Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6, 7, 16

Aaron Hurst et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 7, 16

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 7036–7050, 2024. 3

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7969–7992, 2023. 3

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. *Transactions on Machine Learning Research*, 2025a. 1, 2

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025b. 7

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. VideoSeek: Long-horizon video agent with tool-guided seeking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5465–5475, 2026. 1, 2, 3, 6, 7, 17, 18

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. 1

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems*, volume 38, pp. 168008–168033, 2025. 1, 3

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. DrVideo: Document retrieval based long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946, 2025. 1, 3, 6, 17, 18

Ziqi Pang and Yu-Xiong Wang. MR. Video: MapReduce as an effective principle for long video understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 6640–6672, 2025. 1, 3, 6, 17

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. Introducing Gemini 2.0: Our new AI model for the agentic era. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/, 2024. Accessed 2026-09-25. 6, 7, 16

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055, 2025. 1

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026a. Accessed 2026-09-25. 6, 16

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, 2026b. Accessed 2026-09-25. 6, 7

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599, 2025. 2

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169, 2025. 2

Aaditya Singh et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 6, 7, 16

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, 2025. 2

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The convergence of autonomous driving and large vision-language models. In *Proceedings of the 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4698–4726, 2025. 1

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*, 2024a. 1, 2

Tianyue Wang, Xuying Wu, Yuxiang Ma, Ruiming Liang, Jiaxuan Kang, Yanchao Hao, Zheng Wei, Leigang Qu, Haiyun Guo, and Jinqiao Wang. Routing before looking: Query-adaptive evidence acquisition for long-form video understanding. *arXiv preprint arXiv:2608.20805*, 2026a. 2, 3, 6, 17

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 (ICCV)*, pp. 22958–22967, 2025a. 1, 2, 6, 16

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 (ECCV)*, pp. 58–76. Springer, 2024b. 1, 2, 3, 6, 17, 18

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3283, 2025b. 1, 3, 6, 17, 18

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9088–9099, 2026b. 2

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024. 1, 2, 6, 16

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. VCA: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20168–20179, 2025. 2, 6, 17

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. VideoARM: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24042–24051, 2026. 1, 2, 18

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. VideoExplorer: Advancing long-horizon video understanding via hierarchical orchestration. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6350–6361. ACM, 2026. 2

Z.ai. GLM-5.3-Flash: Frontier intelligence, flash cost. https://z.ai/blog/glm-5.3-flash, 2026. Accessed 2026-09-25. 6, 7

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep Video Discovery: Agentic search with tool use for long-form video understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 89863–89895, 2025. 1, 3, 6, 17, 18

Yaolun Zhang, Ruohui Wang, Jiahao Wang, Yepeng Tang, Xuanyu Zheng, Haonan Duan, Hao Lu, Hanming Deng, and Lewei Lu. EVA: Efficient reinforcement learning for end-to-end video agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12289–12299, 2026. 2, 18

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023. 1

A APPENDIX

The supplementary material provides extended architectural specifications, toolset execution semantics, algorithmic workflows, computational cost accounting, implementation details, extended ablation studies, additional related work, and discussion of limitations to complement the primary manuscript:

- §A.1: **Toolset Specifications and Execution Semantics.** Complete taxonomy, input-output interfaces, and framework versus model action separation for all five tools.
- §A.2: **Complete Algorithmic Workflow.** Step-by-step procedural pseudocode for dual-track retrieval, predictive tier allocation, and bounded iterative verification.
- §A.3: **Cost Structure and Computational Accounting.** Analysis of online question-level inference tokens versus video-level on-demand captioning expenditure.
- §A.4: **Extended Experimental Setup and Implementation Details.** Comprehensive benchmark specifications, task taxonomies, answering backbone endpoint protocols, baseline comparison standards, and low-level retrieval parameters.
- §A.5: **Extended Ablation Studies and Analysis.** Empirical evaluations of initial answer persistence across turns and portability across captioners and backbones.
- §A.6: **Additional Related Work.** Extended discussion of agentic inspection loops and indexed video pipelines.
- §A.7: **Limitations and Future Work.** Discussion of edge-deployment context distillation, continuous temporal grounding, and prospective extensions to streaming video feeds.

A.1 TOOLSET SPECIFICATIONS AND EXECUTION SEMANTICS

To reconcile evidence sufficiency with computational efficiency, ForeSeek establishes a strict architectural division between framework-level evidence assembly and model-level action execution, as detailed in Tab. 6. Standard video agents typically place open-ended search primitives directly in the model's iterative action space, leaving the number and placement of searches to the model. In contrast, ForeSeek executes global evidence acquisition entirely at the framework level before the answering model generates its initial response. Specifically, GLOBALBROWSE uniformly samples F grounding frames across the entire video duration using a standard video decoder, providing a query-independent visual baseline that preserves holistic coverage over quiet intervals. Concurrently, SEMANTICRETRIEVE evaluates cross-modal semantic relevance across visual frames and speech transcripts using off-the-shelf encoders (SigLIP, Whisper, and BGE), outputting temporally dispersed candidate moments without invoking vision-language model generation. Next, BUDGETPROBE conditions the answering model on the smallest book and decodes a few tokens to read its predictive likelihood over the candidate options, allocating the budget $M^{\star}$. Consequently, only two primitives are exposed to the model's action space: INSPECT, which lets the model examine a localized span by fetching frames, descriptions, or transcript lines within bounded turns, and ANSWER, which commits a choice letter and, on the final turn, terminates the episode. This structural decoupling removes open-ended search from the model's action space, replacing it with bounded inspection of a fixed evidence package.

Table 6: Toolset of ForeSeek. The framework executes the first three tools before the first reasoning-and-answer pass. Only INSPECT and ANSWER reside in the model's action space.

Tool	Input	Returns	Backbone
GLOBALBROWSE	count F	uniform grounding frames	video decoder
SEMANTICRETRIEVE	queries, size M	M moments across tracks	SigLIP, Whisper, BGE
BUDGETPROBE	smallest book	allocated budget $M^{\star}$	answering model
INSPECT	span, channels, frame count	localized span evidence	video decoder, text index
ANSWER	choice letter	committed prediction	answering model

me

A.2 Complete Algorithmic Workflow

Algorithm 1 formalizes the end-to-end execution workflow of ForeSeek. The framework begins by executing SemanticRetrieve and GlobalBrowse to obtain the ranked candidate moments $\mathcal{P}$ and the tier-invariant grounding frames $\mathbf{F}$. Next, BudgetProbe reads the model's predictive distribution over the candidate options on the smallest book $\mathcal{B}_{M_{\text{small}}}$ to assign the evidence tier $M^\star$. The answering model is then supplied with the front-loaded package $\tau = \langle \mathbf{F}, \mathcal{B}_{M^\star}, \mathbf{Q} \rangle$. On each turn the model commits a provisional answer $\mathbf{Y}_t$ and names a span a_t to inspect, mandatory on the first turn and optional thereafter. A span overlapping what has been shown by more than half is refused once and the model asked to re-target; a request that would push the fetched frames past a fraction ρ of the video ends the loop. Otherwise the localized evidence v_t is fetched and appended to τ, for up to H requests, before the final answer $\mathbf{Y}$ is returned.

Algorithm 1 ForeSeek answering workflow

Require: question $\mathbf{Q}$; video $\mathbf{X}$; answering model p_θ; frame count F; tier map $M(\cdot)$ with thresholds $\gamma_1 < \gamma_2$; request limit H; frame-budget fraction ρ
Ensure: answer $\mathbf{Y}$

$\bar{s},\ \mathcal{P} \leftarrow$ SemanticRetrieve$(\mathbf{X}, \mathbf{Q})$ ▷ dual-track scoring in Sec. 3.2
$\mathbf{F} \leftarrow$ GlobalBrowse$(\mathbf{X}, F)$ ▷ uniform frames, fixed across tiers
$M^\star \leftarrow$ BudgetProbe$(\mathbf{F}, \mathcal{B}_{M_{\text{small}}}, \mathbf{Q})$ ▷ few-token probe, Eqs. 5 and 6
$\tau \leftarrow \langle \mathbf{F}, \mathcal{B}_{M^\star}, \mathbf{Q} \rangle$; $\mathcal{S} \leftarrow \emptyset$ ▷ front-load allocated package
for $t = 1$ **to** $H + 1$ **do**
  $(z_t, \mathbf{Y}_t, a_t) \leftarrow p_\theta(\tau)$ ▷ reason, commit answer, request span
  $\mathbf{Y} \leftarrow \mathbf{Y}_t$
  if $t = 1$ **and** $a_t = \varnothing$ **then**
    $a_t \leftarrow p_\theta(\tau, I_{\text{req}})$ ▷ mandatory first request
  end if
  if $a_t = \varnothing$ **or** $t = H + 1$ **then**
    break ▷ model settled, or request limit reached
  end if
  if Overlap$(a_t, \mathcal{S}) > \frac{1}{2}$ **then**
    $a_t \leftarrow p_\theta(\tau, I_{\text{rej}})$ ▷ refused once; model re-targets
    if $a_t = \varnothing$ **or** Overlap$(a_t, \mathcal{S}) > \frac{1}{2}$ **then**
      break
    end if
  end if
  if FetchedSpan$(\mathcal{S}) + |a_t| > \rho\,|\mathbf{X}|$ **then**
    break ▷ frame budget exhausted
  end if
  $v_t \leftarrow$ Inspect$(\mathbf{X}, a_t)$; $\mathcal{S} \leftarrow \mathcal{S} \cup \{\text{span}(a_t)\}$
  $\tau \leftarrow \tau \cup \langle z_t, \mathbf{Y}_t, a_t, v_t \rangle$
end for
return $\mathbf{Y}$

A.3 Cost Structure and Accounting

A rigorous comparison of long-video systems requires distinguishing two distinct computational costs that are frequently confused in the literature. The first quantity is the online question-level inference cost, which encompasses the tokens processed by the answering model during inference. In ForeSeek, this corresponds to the invariant F visual frames, the $M^\star$ retrieved multimodal moments, and any localized frames fetched during bounded inspection. The second quantity is the video-level description cost, representing the vision language model passes required to generate visual clip descriptions. Indexed pipelines typically construct an offline dense index over the entire video before any query is received, hiding a heavy offline captioning cost behind an apparently compact online token count.

ForeSeek dramatically reduces the video-level description cost without sacrificing retrieval accuracy. Because the dual-track retriever ranks candidate moments using frozen image and sentence encoders

operating on raw frames and speech transcripts, the framework does not require visual descriptions to exist in advance. Instead, clip descriptions are generated on demand only for the sparse candidate moments returned by the retriever. Under this on-demand formulation, our predictive tier allocation translates directly into computational savings for video captioning. Queries assigned to the smaller tier require captioning only ten moments, saving considerable vision language model calls compared with uniform high-tier configurations. When multiple questions query the same video, descriptions of overlapping intervals are cached and shared, amortizing the description cost across queries. We evaluate the quantitative impact of this cost structure in Sec. 4.2.

A.4 Extended Experimental Setup and Implementation Details

Benchmark specifications and task taxonomies. We evaluate ForeSeek across four benchmarks covering both broad long video understanding and complex temporal reasoning:

- **LVBench** (Wang et al., 2025a) comprises 1,549 multiple-choice questions across 103 videos with an average duration exceeding one hour. It evaluates six core long-video competencies: entity recognition, event understanding, key information retrieval, temporal grounding, reasoning, and summarization.
- **Video-MME** (Fu et al., 2025) benchmarks multimodal video reasoning across diverse temporal durations. Following standard long-video evaluation settings, we benchmark on its long subset consisting of 900 questions over 300 videos lasting between 30 and 60 minutes.
- **LongVideoBench** (Wu et al., 2024) assesses extended context understanding over diverse video domains. We evaluate on its long validation split comprising 564 questions over 188 videos ranging from 15 to 60 minutes.
- **Video-Holmes** (Cheng et al., 2025) evaluates multi-clue synthesis and temporal deduction over 1,837 questions distributed across 270 suspense short films. Unlike conventional benchmarks where superficial temporal sweeping often suffices, Video-Holmes requires multi-step deduction over dispersed and subtle clues across visual and auditory streams. It benchmarks seven fine-grained reasoning categories:
  - **Social Reasoning**, denoted SR: Inferring interpersonal dynamics, character motives, social hierarchy, and tacit personal relationships.
  - **Intention and Motive Chaining**, denoted IMC: Decoding multi-step psychological motives, deceptive goals, and hidden agendas driving character actions.
  - **Temporal Causal Inference**, denoted TCI: Identifying chronological cause-and-effect chains connecting non-adjacent events across the narrative arc.
  - **Timeline Analysis**, denoted TA: Reconstructing global chronologies, aligning parallel subplots, and verifying temporal markers across disjoint scenes.
  - **Multimodal Hint Reasoning**, denoted MHR: Grounding reasoning in subtle visual props, background audio cues, and symbolic foreshadowing across modalities.
  - **Physical Anomaly Reasoning**, denoted PAR: Spotting physical contradictions, unnatural interactions, and fabricated physical evidence that violate physical laws.
  - **Core Theme Inference**, denoted CTI: Integrating fragmented narrative evidence to deduce high-level thematic conclusions and central plot resolutions.

Confidence probe and tier thresholds. For confidence-driven budget allocation, each backbone reads its predictive probability over the candidate option tokens via the few-token probe on the smallest book $\mathcal{B}_{M_{\text{small}}}$, giving the confidence c. Thresholds $\gamma_1 = 0.6$ and $\gamma_2 = 0.9$ route queries to the small tier with $M = 10$, the mid tier with $M = 20$, or the large tier with $M = 30$.

Baseline systems and comparison standards. We compare ForeSeek against two representative families in the literature:

- **General-purpose VLMs:** Frontier vision-language foundation models operating under standard direct answering with uniform frame sampling, including Qwen2.5-VL-72B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), Qwen3.5-9B (Qwen Team, 2026a), Gemini-1.5-Pro (Gemini Team et al., 2024), Gemini 2.0 Flash (Pichai et al., 2024), GPT-4o (Hurst et al., 2024), GPT-5 (Singh et al., 2025), alongside our direct-answering backbones Qwen3.8-27B and GLM-5.3-Flash evaluated with 96 or 384 uniform frames.

- **Agentic video systems:** State-of-the-art interactive agents that iteratively explore long videos, including VideoAgent (Wang et al., 2024b), VideoTree (Wang et al., 2025b), DrVideo (Ma et al., 2025), VCA (Yang et al., 2025), Mr.Video (Pang & Wang, 2025), DVD (Zhang et al., 2025), VideoSeek (Lin et al., 2026), and Route2Look (Wang et al., 2026a).

In long-form video understanding, speech dialogue provides indispensable factual anchors. Prior agent benchmarks report substantial performance gaps depending on whether human-curated subtitles are supplied. To guarantee fair, consistent, and realistic evaluation across all benchmarks, ForeSeek does not rely on pre-packaged subtitle files distributed with benchmarks such as Video-MME or LongVideoBench. Instead, all videos across all four benchmarks are transcribed directly from raw audio tracks using Whisper. When comparing against published results of prior agent methods, we report their with-subtitle configurations wherever available, ensuring strict information parity since ForeSeek similarly utilizes spoken content.

Retrieval pipeline and temporal curve smoothing. Transcript segments are merged into runs of at most 30 s, broken at silences longer than 5 s, where w_t denotes the run containing second t. The visual relevance curve is standardized over seconds and smoothed with a 5 s sliding window. The speech relevance curve is standardized over spoken runs, with seconds falling outside speech segments set to zero. Non-maximum suppression enforces a minimum temporal separation of 5 s across all candidate book sizes. Consequently, each evidence book forms a strict prefix of the global ranking, guaranteeing strictly nested evidence packages as defined in Eq. 4. Candidate clips are set to duration $L = 5$ s, ensuring selected seconds correspond to distinct intervals. Under scene-based video segmentations where scene durations exceed 5 s, any scene containing multiple selected seconds is included in the evidence book once.

Visual frame allocation and sampling. The grounding frames $\mathbf{F}$ are sampled on a uniform grid over the full video duration without reference to the question, with $F = 96$ ($F = 48$ for Qwen3.8-27B on LongVideoBench); videos shorter than $2F$ seconds receive one frame every two seconds, with a floor of 64. Frames are resized to a short side of 256 pixels and tiled eight to an image in a 2×4 grid, each cell stamped with its timestamp, so that a frame costs about 114 visual tokens. Frames returned by INSPECT are sampled uniformly within the requested span, which is clipped to 90 s for frames and 600 s for text; the model chooses the count between 8 and 128, defaulting to 48.

A.5 Extended Ablation Studies and Analysis

Persistence of the first-turn answer under bounded inspection. Tab. 7 traces how often the answer committed on the first turn survives the inspection turns that follow, across four benchmarks and two answering backbones. On 89.4%–98.9% of questions the final answer is the one committed on the first turn. Restricting to questions answered incorrectly on the first turn, 71.7%–97.5% keep that answer through the rest of the trajectory: 72.3% for Qwen3.8-27B and 79.4% for GLM-5.3-Flash on LVBench, rising to 97.5% for GLM-5.3-Flash on Video-Holmes, where inspection almost never changes the outcome. Within ForeSeek, then, the first-turn answer carries most of the weight, so the evidence that decides most questions has to be present before that answer is written; this is the case the Pre-Answer Evidence Allocator is built for, with bounded inspection acting as a second-order correction for the remaining 2.5%–28.3% of initial errors that are revised. These figures are measured on ForeSeek's own trajectories, in which inspection is confined to short local spans.

Table 7: Persistence of the first-turn answer on ForeSeek's own trajectories. Initial is the answer committed on the first turn from the front-loaded package $\mathbf{E}_{M^\star}$; final is the answer returned after bounded local inspection. First = final: share of all questions whose final answer equals the initial one. Unrevised initial errors: share of initially incorrect answers that are unchanged at the end of the trajectory; a revised answer may or may not be correct. Both metrics are percentages.

Dataset	Backbone	First = Final	Unrevised initial errors
LVBench	Qwen3.8-27B	90.5	72.3
LVBench	GLM-5.3-Flash	94.2	79.4
Video-MME	Qwen3.8-27B	96.4	87.8
LongVideoBench	Qwen3.8-27B	89.4	71.7
Video-Holmes	GLM-5.3-Flash	98.9	97.5

Compatibility across captioners and answering backbones. Tab. 8 evaluates compatibility across captioners and backbones on LVBench. Gemini-2.0-Flash delivers roughly half the text volume of GPT-4.1, while frame allocations remain virtually unchanged at 116.8–119.0 for Qwen3.8-27B and 109.6–109.8 for GLM-5.3-Flash. This text reduction incurs a modest drop of 3.2 percentage points on Qwen3.8-27B from 75.7% to 72.5%, and only 1.3 percentage points on GLM-5.3-Flash from 78.5% to 77.2%. GLM-5.3-Flash maintains superior performance under both captioners, peaking at 78.5%. These results show that ForeSeek generalizes across description styles while preserving backbone rankings and stable frame consumption.

Table 8: **Compatibility with different captioners and answering backbones on LVBench.** All rows share one retriever, one ladder of packet sizes and one answering protocol. Text halves across captioners while #Frames does not, so accuracy is comparable between backbones but not between captioners.

Captioner	Backbone	Acc.	Text (k)	Frames
Gemini-2.0-Flash	Qwen3.8-27B	72.5	14.5	119.0
GPT-4.1	Qwen3.8-27B	75.7	29.0	116.8
Gemini-2.0-Flash	GLM-5.3-Flash	77.2	14.0	109.6
GPT-4.1	GLM-5.3-Flash	**78.5**	27.3	109.8

A.6 Additional Related Work

Agentic inspection. VideoAgent (Wang et al., 2024b) uses a language model as a central agent that iteratively gathers the information a question needs, with vision–language models as tools to retrieve and describe frames. VideoSeek (Lin et al., 2026) follows the logic flow of the video in a think–act–observe loop, moving from a coarse overview to skimming candidate intervals and focusing on short clips. EVA (Zhang et al., 2026) plans before it perceives through summary, plan, action and reflection steps, and is trained end to end with supervised fine-tuning and reinforcement learning. VideoARM (Yin et al., 2026) additionally keeps a hierarchical multimodal memory, built on the fly as a controller invokes tools from coarse to fine in a loop of observing, thinking, acting and memorizing.

Indexed pipelines. DrVideo (Ma et al., 2025) converts a long video into a coarse text document, retrieves key frames from it, and runs an agent loop that augments the document until enough question-related information is gathered. VideoTree (Wang et al., 2025b) iteratively refines keyframes by their relevance to the query and organizes them into a coarse-to-fine tree over which a language model reasons. DVD (Zhang et al., 2025) captions every short clip with a vision–language model, stores the captions, embeddings and frames in a multi-granular database, and lets a reasoning model search it with browsing, clip search and frame inspection tools.

A.7 Limitations and Future Work

ForeSeek trades thousands of raw visual frames for compact, channel-aligned textual evidence, bypassing the quadratic memory bottlenecks of dense video attention while preserving essential multimodal clues. In deployment regimes governed by strict memory constraints, such as resource-limited edge hardware, accumulating extensive speech transcripts and moment descriptions across multiple intervals still introduces a noticeable textual context footprint. A promising direction for future research is exploring adaptive cross-modal token distillation and extractive transcript summarization to further compress the evidence book without sacrificing decisive factual details.

Additionally, ForeSeek operates on segmented temporal intervals and pre-recorded videos where audio tracks and visual candidates are indexed upfront. Extending pre-answer calibration to sub-second continuous action grounding and real-time streaming video feeds where future context is unavailable represents an important avenue for future exploration.