HIERASEEK: LEARNING TO SEARCH VISUAL HIERARCHIES FOR LONG-VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have achieved remarkable progress in video understanding, yet long-video reasoning remains challenging because task-relevant evidence is sparse and easily obscured by redundant visual content. Existing methods typically compress long videos into fixed representations or iteratively search over flat temporal sequences, which can limit effective evidence discovery. We propose **HieraSeek**, a hierarchical framework that constructs a reusable multi-granularity visual representation directly within the ViT encoder through progressive frame pruning and temporal merging, and exposes this representation as the search space for multi-round reasoning. This shared latent structure tightly couples representation construction with evidence acquisition, avoiding a separate retrieval or memory pipeline during reasoning. A learned policy iteratively refines or relocates its temporal focus, while reusing intermediate visual states to reduce repeated computation. We further optimize the search policy with an evidence-seeking reinforcement learning objective. Across six long-video understanding and reasoning benchmarks, HieraSeek achieves the best performance on all reported metrics, outperforming strong multi-round baselines under the same visual budget while maintaining competitive inference efficiency.

1 INTRODUCTION

Recent advances in Multimodal Large Language Models (MLLMs) (Bai et al., 2025a; Wang et al., 2025b; Feng et al., 2025; Comanici et al., 2025) have substantially improved video understanding across tasks such as video question answering (Antol et al., 2015; Kim et al., 2025), temporal grounding (Gao et al., 2017; Ren et al., 2024; Lin et al., 2023), and complex video reasoning (Zhu et al., 2026). However, reasoning over long videos remains challenging: task-relevant evidence is often sparse and temporally dispersed, while much of the visual content is redundant or irrelevant to a given query. Consequently, the challenge is no longer only how to fit more video content into a limited context, but also *how visual information should be organized and selectively accessed during reasoning*.

Existing approaches broadly follow two complementary directions. Compression-based methods reduce the visual context through frame selection, visual token pruning, or adaptive spatiotemporal compression (Zhang et al., 2025; Huang et al., 2025; Shen et al., 2024), but the resulting representation is typically fixed before reasoning and may irreversibly discard useful evidence. More recent agentic approaches, instead, enable MLLMs to iteratively inspect additional temporal regions, allowing the model to refine or correct its visual focus during reasoning (Ding et al., 2025; Yang et al., 2025; He et al., 2026). Although effective, such methods commonly treat evidence acquisition and visual representation as separate stages: reasoning repeatedly queries clips from the original temporal sequence, while the underlying video representation provides limited structure for searching across temporal scales.

Recent studies have further explored hierarchical and multi-granularity organization of long videos to support coarse-to-fine evidence access and reasoning (Wang et al., 2025c; Qiu et al., 2026; Yin et al., 2026a). These works demonstrate the value of structured temporal organization for long-video understanding, but their hierarchies are typically constructed as external retrieval or memory structures over already processed video content. As illustrated in Fig. 1, existing paradigms either compress visual information into a fixed representation before reasoning or iteratively search over

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.

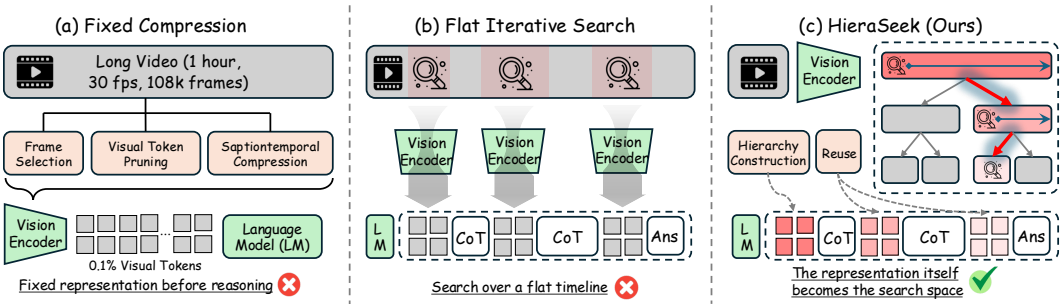


Figure 1: **Conceptual comparison of long-video reasoning paradigms.** (a) Compression-based methods compact visual information before reasoning, resulting in a fixed representation that may discard task-relevant evidence. (b) Agentic methods enable iterative evidence acquisition, but typically search over a flat temporal sequence and repeatedly process queried clips. (c) HieraSeek constructs a searchable visual hierarchy directly within the vision encoder, allowing multi-round reasoning to progressively access and reuse visual representations across temporal granularities.

the original temporal sequence, whereas our goal is to make the visual representation itself directly searchable during reasoning. This motivates a different question: *Can the visual representation itself serve as a searchable hierarchy, so that representation construction and multi-round reasoning operate over the same latent structure?*

In this work, we propose **HieraSeek**, a hierarchical framework for long-video reasoning that couples multi-granularity visual representation with learned evidence search. Our design is motivated by two observations. First, long videos contain substantial local redundancy, making it unnecessary to preserve every frame with equal fidelity. Second, frame representations at different depths of a Vision Transformer (ViT) exhibit different levels of temporal similarity, suggesting that the visual encoder itself provides a natural basis for organizing video content at multiple granularities. Based on these observations, we introduce progressive frame pruning and temporal merging across early ViT blocks, producing a tree-structured visual hierarchy in which temporally contiguous regions are represented at progressively coarser scales.

Crucially, this hierarchy is not used merely as a compressed representation. We expose it directly to the reasoning process as a structured search space. At each reasoning round, the model can refine its current temporal focus or relocate to another region, retrieving finer-grained visual evidence from the corresponding hierarchy nodes. Because the hierarchy is constructed from intermediate ViT states, previously computed representations can be reused across rounds, reducing repeated computation during multi-step search. We further train the model to perform this search with reinforcement learning, using an evidence-seeking reward that encourages the policy to localize supporting evidence while avoiding unnecessary exploration. In this way, HieraSeek jointly addresses how long-video information is represented and how it is progressively accessed during reasoning.

We evaluated HieraSeek on six long-video understanding and reasoning benchmarks. The results show consistent improvements over the base model and strong performance compared with recent single-round and multi-round reasoning approaches. Further analyses validate its effectiveness and efficiency. In summary, our main contributions are as follows.

I. We introduce a multi-granularity visual hierarchy constructed directly within the ViT encoder through progressive frame pruning and temporal merging, providing a compact and structured representation of long videos across temporal scales.

II. We formulate long-video reasoning as learned search over this hierarchy, enabling the model to progressively refine or relocate its temporal focus, and introduce an evidence-seeking reinforcement-learning objective to optimize the search behavior.

2 Related Work

Long-Video Representation and Compression. Scaling MLLMs to long videos requires handling substantial temporal redundancy under limited visual-context budgets. One line of work extends the amount of video content that can be processed through long-context modeling and efficient

sequence processing (Zhang et al., 2024b; Chen et al., 2024b). Another line reduces the visual context through frame selection, visual token pruning or merging, and adaptive spatiotemporal compression (Zhang et al., 2025; Huang et al., 2025; Shen et al., 2024). Memory-based approaches further summarize or store long-video content in compact representations that can be retrieved at inference time (Zhang et al., 2024a; Man et al., 2025). These methods substantially improve the scalability of video MLLMs, but the resulting representations are generally constructed primarily to reduce or retrieve context. In contrast, our focus is on constructing a visual representation that also serves directly as the search space for subsequent multi-round reasoning.

Iterative Long-Video Reasoning. Recent work increasingly treats long-video understanding as an active evidence-acquisition process rather than one-shot inference over uniformly sampled frames. VideoChat-A1 (Wang et al., 2026) progressively narrows temporal regions through chain-of-shot reasoning, while LongVT (Yang et al., 2025) and VideoZoomer (Ding et al., 2025) interleave reasoning with temporal cropping or zooming to acquire finer-grained evidence. FrameThinker (He et al., 2026) learns multi-turn frame spotlighting through supervised and reinforcement learning, and Video-o3 (Zeng et al., 2026) supports native interleaved clue seeking with adaptive temporal inspection. Related agentic approaches further explore active perception, tool-augmented reasoning, and adaptive stopping for long-video understanding. These methods demonstrate the effectiveness of iterative evidence search, but typically perform evidence acquisition over frames or temporal clips selected from the video sequence. Our work instead performs learned search over a reusable multi-granularity visual representation constructed within the encoder.

Hierarchical Video Organization. Hierarchical representations have recently emerged as a complementary approach for organizing long-video content across temporal granularities. VideoTree (Wang et al., 2025c) constructs a query-adaptive tree through iterative breadth and depth expansion for coarse-to-fine retrieval. LongVideo-R1 (Qiu et al., 2026) starts from high-level visual summaries and progressively refines its temporal focus through learned navigation. More recent systems organize long videos into hierarchical or multi-scale memories and combine them with agentic retrieval, including VideoARM (Yin et al., 2026b), HAVEN (Yin et al., 2026a), and WorldMM (Yeo et al., 2026). These approaches highlight the value of multi-granularity organization for long-video reasoning. Our method differs in where the hierarchy is formed and how it is used: we construct it progressively from intermediate representations within the ViT encoder and expose the same latent structure directly to a learned multi-round search policy.

3 METHODOLOGY

3.1 OVERVIEW: SEARCHABLE VISUAL HIERARCHY

Given a long video and a question q, our goal is to construct a compact visual representation that can be progressively accessed during reasoning, rather than compressed once and consumed as a fixed context.

As illustrated in Fig. 2, HieraSeek organizes the video into a multi-granularity visual hierarchy $\mathcal{H}$ directly within the ViT encoder. Each hierarchy node corresponds to a contiguous temporal region and is associated with intermediate visual states at a particular encoder depth. Nodes at different levels therefore provide representations of the video at different temporal granularities.

Rather than treating $\mathcal{H}$ only as a compressed representation, we expose it as the search space for multi-round reasoning. At each round, the model can refine its current temporal focus, relocate to another region, or terminate the search and produce an answer. The hierarchy is constructed once during the initial encoding stage and reused throughout subsequent search rounds.

3.2 PROGRESSIVE REPRESENTATION CONSTRUCTION

Our construction is motivated by two empirical observations: temporally nearby frames exhibit substantial local redundancy, and the degree of inter-frame similarity varies across ViT depths (see Appendix A). Rather than compressing the video once at a fixed representation level, we progressively reduce redundant frames across early ViT blocks while preserving representative content over contiguous temporal regions.

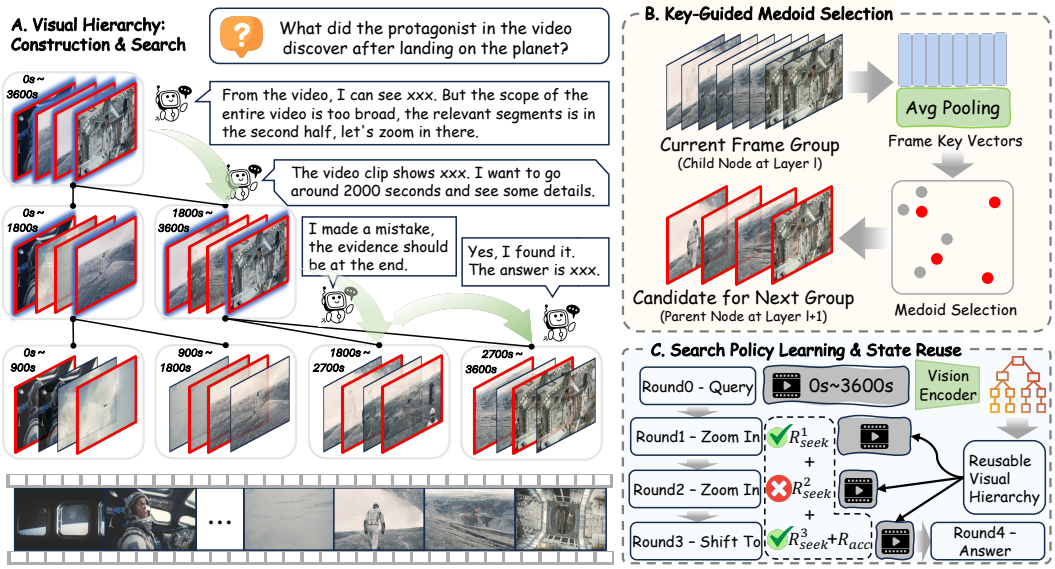


Figure 2: Overview of **HieraSeek**. (A) A long video is organized into a multi-granularity visual hierarchy, which the model iteratively searches using `zoom_in`, `shift_to` and `answer` actions to localize task-relevant evidence. (B) Each hierarchy node is compressed through key-guided medoid selection, which retains representative frames for progressive hierarchy construction. (C) The search policy is optimized with supervised trajectories and an evidence-seeking reinforcement-learning objective, while the constructed hierarchy is reused across reasoning rounds.

Frame-Level Key Representation. Let the densely sampled video be $V = \{f_1, f_2, \ldots, f_N\}$, where each frame f_t contains M visual tokens. At the ℓ-th ViT block, we denote the hidden representation of the s-th token in frame t as $h_{t,s}^{(\ell)} \in \mathbb{R}^d$. Its attention key is

$$k_{t,s}^{(\ell)} = h_{t,s}^{(\ell)} W_K^{(\ell)}, \tag{1}$$

where $W_K^{(\ell)}$ is the key projection matrix. We obtain a frame-level representation by spatially averaging the key vectors,

$$F_t^{(\ell)} = \frac{1}{M} \sum_{s=1}^{M} k_{t,s}^{(\ell)}. \tag{2}$$

These frame-level key representations are used to measure redundancy within local temporal groups.

Key-Guided Medoid Selection. At level ℓ, we partition the current frame sequence into non-overlapping groups of n temporally consecutive frames. For each group $g_{\ell,j}$, we select k representative frames from the original group. For two frames a and b within the same group, we define

$$D_j^{(\ell)}(a,b) = 1 - \cos\left(F_{j,a}^{(\ell)}, F_{j,b}^{(\ell)}\right) + \alpha_{\text{temp}} \frac{|a-b|}{n-1}, \tag{3}$$

where the first term measures feature dissimilarity in the ViT key space, while the second term incorporates temporal distance to encourage coverage across the group.

We formulate frame selection as a medoid-style subset selection problem:

$$S_{\ell,j}^* = \arg \min_{\substack{S \subseteq \{1,\ldots,n\} \\ |S|=k}} \sum_{a=1}^{n} \min_{b \in S} D_j^{(\ell)}(a,b). \tag{4}$$

The selected elements are actual input frames and jointly provide a compact representation of the visual and temporal variation within the group. Since solving Eq. 4 exactly is combinatorial, we use a greedy approximation. The full procedure is provided in Appendix B.

Progressive Temporal Merging. We set $k = \lfloor n/2 \rfloor$. After selection, the retained frames from two temporally adjacent groups are concatenated in their original temporal order to construct the group at the next level:

$$g_{\ell+1,m} = \mathrm{Concat}_{\mathrm{time}}\left(g_{\ell,2m-1}[S_{\ell,2m-1}], g_{\ell,2m}[S_{\ell,2m}]\right). \tag{5}$$

Because each child retains $n/2$ frames, the resulting parent group again contains n frames and can be processed by the subsequent ViT block using the same procedure. Moreover, because only temporally adjacent groups are merged, every group continues to correspond to a contiguous interval in the original video.

Repeating this selection-and-merging process progressively reduces the number of retained frames while constructing representations over increasingly larger temporal spans. Once the number of retained frames reaches the budget, we stop pruning and forward the remaining visual tokens through the rest of the ViT encoder to obtain the global video representation. Intermediate states produced during this process are cached for subsequent multi-round search, as described in Sec. 3.3.

3.3 Multi-Round Evidence Search

Given the visual representation constructed above, HieraSeek performs long-video reasoning as an iterative evidence-search process. At reasoning round t, the model maintains a temporal focus $I_t = [l_t, r_t]$ and selects an action from

$$\mathcal{A} = \{\texttt{zoom_in}, \texttt{shift_to}, \texttt{answer}\}.$$

The initial interval I_0 covers the full video. The `zoom_in` action selects a stricter sub-interval of the current region when finer-grained evidence is needed, while `shift_to` relocates the search to another temporal region when the current focus is insufficient or incorrect. The `answer` action terminates the search and produces the final response once sufficient evidence has been collected.

Multi-Granularity Evidence Access. For either `zoom_in` or `shift_to`, the model additionally predicts a target temporal interval I_{t+1}. We map this interval to a compact set of hierarchy nodes,

$$\mathcal{C}_{t+1} = \mathrm{Cover}_{\mathcal{H}}(I_{t+1}; P),$$

where $\mathcal{H}$ denotes the visual hierarchy constructed in Sec. 3.2, and P is a bounded padding budget used to accommodate interval boundaries that do not align exactly with the hierarchy partitions.

The resulting node set provides the highest available temporal granularity that covers the requested region, allowing the model to access the target evidence without reverting to the original frame sequence. The interval-to-node mapping procedure is detailed in Appendix C.

Intermediate State Reuse. Each retrieved node is associated with the intermediate visual states produced at the ViT block where that node was constructed. Rather than re-encoding the underlying frames from the input, we reuse these cached states and resume computation from the subsequent ViT blocks without further pruning. The resulting features form the local visual representation L_t for the current reasoning round. This reuse avoids repeatedly processing the same visual content across search rounds.

Reasoning Context. At each round, the language model conditions on both the global video representation and the visual evidence retrieved for the current temporal focus. We write the reasoning context as

$$X_t = [G, L_t, q, H_{<t}],$$

where G denotes the global video representation obtained in the first reasoning round, L_t the current local visual evidence, q the input question, and $H_{<t}$ the reasoning and action history from previous rounds. For efficiency, local visual tokens from earlier rounds are not retained explicitly. Instead, their relevant information is carried forward through $H_{<t}$. This combination preserves global context while allowing the model to progressively inspect task-relevant regions at finer or alternative temporal locations.

3.4 Learning the Search Policy

We train the search policy in two stages: supervised fine-tuning on multi-round search trajectories, followed by reinforcement learning to further optimize its evidence-seeking behavior.

Supervised Fine-Tuning (SFT). We construct multi-round trajectories that interleave reasoning with the actions defined in Sec. 3.3. At each round, the model observes the current visual context and the preceding trajectory, and predicts either a search action together with a target temporal interval or the final answer. The trajectories are generated by a teacher model and automatically filtered for answer correctness, action validity, and consistency with annotated evidence intervals. SFT on these trajectories provides an initial policy for multi-round evidence localization.

Evidence-Seeking Reinforcement Learning (RL). We further optimize the supervised policy with an evidence-seeking reward that encourages the model to localize supporting evidence while avoiding unnecessary search.

Let I_{t+1} denote the temporal interval reached after an intermediate search action at round t, and let I^* be the annotated evidence interval. We measure evidence coverage as

$$\rho_t^{\mathrm{cov}} = \frac{|I_{t+1} \cap I^*|}{|I^*|}, \tag{6}$$

and define a reduction ratio ρ_t^{red} that measures how much the search region is narrowed by the action. The exact definitions for `zoom_in` and `shift_to` are provided in Appendix D.

For each intermediate search action, we assign

$$q_t = \mathbb{1}[\mathrm{valid}(a_t)]\, \mathbb{1}[\rho_t^{\mathrm{cov}} \geq \eta]\, \rho_t^{\mathrm{cov}} \left(1 - \rho_t^{\mathrm{red}}\right), \tag{7}$$

which favors actions that retain sufficient evidence while reducing the search scope.

The evidence-seeking reward is accumulated over search actions only:

$$R_{\mathrm{seek}} = c(T) \sum_{t=0}^{T-1} \gamma^t q_t, \qquad c(T) = \frac{1}{2}\left(1 + \cos\frac{\pi T}{T_{\max}}\right), \tag{8}$$

where T denotes the number of reasoning rounds before the final answer, γ is a discount factor, and $c(T)$ discourages unnecessarily long search trajectories.

The final answer is evaluated separately using the task-level accuracy reward R_{acc}, and the overall reward is

$$R = R_{\mathrm{acc}} + \lambda_{\mathrm{seek}} R_{\mathrm{seek}}. \tag{9}$$

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate HieraSeek on six benchmarks covering both long-video understanding and long-video reasoning. For general long-video understanding, we use MLVU (Zhou et al., 2025), LongVideoBench (Wu et al., 2024), Video-MME (Fu et al., 2025), and LVBench (Wang et al., 2025a). We further evaluate on VideoMMLU (Song et al., 2025) and Video-MMMU (Hu et al., 2025), which place greater emphasis on knowledge-intensive and multi-step reasoning over long videos. We follow the official evaluation protocols and report accuracy on all benchmarks.

Baselines. We compare HieraSeek with representative long-video reasoning models under single-round and multi-round settings. *Single-round reasoning multimodal models* include Qwen2.5-VL (Bai et al., 2025b), Video-LLaVA (Lin et al., 2024), LLaVA-OneVision (Li et al., 2024), LongVU (Shen et al., 2024), Video-MTR (Xie et al., 2025), Video-R1 (Feng et al., 2025), and VideoTree (Wang et al., 2025c). Among them, VideoTree is particularly relevant to our setting, as it constructs an external hierarchical video representation for query-guided retrieval before final reasoning. *Multi-round reasoning multimodal models* include Conan (Ouyang et al., 2025), LongVT (Yang et al., 2025), VideoZoomer (Ding et al., 2025), and FrameThinker (He et al., 2026). These methods cover both conventional long-video MLLMs and recent approaches that progressively acquire task-relevant visual evidence during reasoning.

Table 1: Comparison with existing methods on long-video understanding and long-video reasoning benchmarks. * indicates evaluation under a shared 128-frame visual budget. **Bold** and underline indicate the best and second-best results, respectively.

Methods	MLVU	LongVideoBench	VideoMME		LVBench	VideoMMLU	VideoMMMU
	dev	*val*	*overall*	*long*	*avg*	*quiz*	*avg*
Single-round reasoning multimodal models							
Qwen2.5-VL*	58.4	51.0	63.6	54.0	36.9	62.0	48.3
Video-LLaVA	36.2	37.6	39.9	–	–	–	–
LLaVA-OneVision	64.7	56.4	58.3	46.7	–	33.4	33.9
LongVU	65.4	–	60.6	–	–	–	–
Video-MTR	–	–	59.0	51.0	–	–	–
Video-R1*	64.9	52.2	61.1	51.3	38.7	61.3	49.8
VideoTree*	62.1	52.4	62.5	53.0	39.0	61.9	49.3
Multi-round reasoning multimodal models							
Conan	63.4	56.6	60.5	–	39.2	–	–
LongVT	–	–	64.3	–	<u>41.3</u>	–	45.4
VideoZoomer*	**68.9**	57.5	<u>65.2</u>	<u>55.7</u>	<u>41.3</u>	<u>67.9</u>	<u>52.2</u>
FrameThinker*	59.1	<u>57.6</u>	64.3	54.0	39.6	63.2	50.1
HieraSeek*	**68.9**	**59.2**	**66.4**	**57.2**	**42.1**	**68.0**	**52.4**

Implementation Details. We use Qwen2.5-VL-7B-Instruct as the base model. For SFT, we construct multi-round search trajectories from VideoSIAH (Yang et al., 2025), CG-Bench (Chen et al., 2024a). The trajectories are synthesized using Gemini 2.5 Pro (Comanici et al., 2025), resulting in 10k training trajectories. We further optimize the SFT model with RL using Dr. GRPO (Liu et al., 2025) on CG-Bench and LongVideoReason (Chen et al., 2025). Detailed data construction and optimization settings are provided in Appendix E.

Unless otherwise specified, the temporal group size is set to $n = 32$, and progressive frame selection is applied at six early ViT blocks. The maximum number of search rounds is $T_{\max} = 8$, and the padding budget for temporal interval mapping is $P = 128$. For the evidence-seeking reward, we set $\gamma = 0.8$, $\eta = 0.75$, and $\lambda_{\text{seek}} = 0.5$.

Evaluation Protocol. We impose a shared cumulative visual budget across all reasoning rounds rather than assigning an independent budget to each round. Specifically, the total visual access is capped at the equivalent of 128 input frames, preventing multi-round methods from receiving additional visual information simply by performing more search steps. Qwen2.5-VL applies a temporal merging factor of two in the visual encoder; therefore, 128 input frames correspond to visual tokens with a temporal dimension equivalent to 64 frames at the LLM input.

4.2 Main Results

Table 1 compares HieraSeek with representative single-round and multi-round approaches across six long-video understanding and reasoning benchmarks. Overall, HieraSeek achieves the best or tied-best performance on all reported metrics, demonstrating consistent gains across both long-video understanding and reasoning tasks.

Comparison with Iterative Evidence Search. Compared with VideoZoomer, a strong prior multi-round baseline evaluated under the same visual budget, HieraSeek improves LongVideoBench, Video-MME overall, Video-MME long, and LVBench by 1.7, 1.2, 1.5, and 0.8 points, respectively, while matching its performance on MLVU. It further improves VideoMMLU

Table 2: Ablation on representation structure and policy learning.

Method	MLVU	LongVideoBench	VideoMME	LVBench	VideoMMLU	VideoMMMU
Base	58.4	51.0	63.6	36.9	62.0	48.3
SFT (Flat)	63.2	53.9	64.3	38.1	64.2	48.5
SFT (Hierarchy)	66.1	56.6	65.6	39.8	65.9	50.6
SFT+RL (Flat)	66.2	55.3	66.0	40.3	66.2	50.2
SFT+RL (Hierarchy)	**68.9**	**59.2**	**66.4**	**42.1**	**68.0**	**52.4**

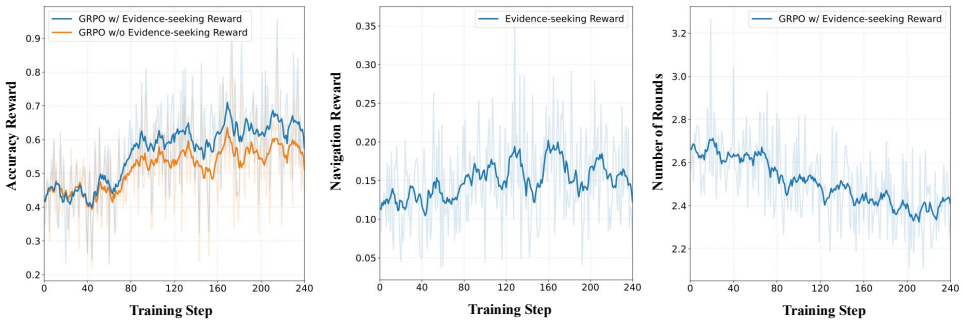


Figure 3: Training dynamics during reinforcement learning. Left: average accuracy reward over training steps for HieraSeek with and without evidence-seeking reward. Middle: average evidence-seeking reward over training steps. Right: average number of reasoning rounds over training steps.

and VideoMMMU by 0.1 and 0.2 points. HieraSeek also consistently outperforms FrameThinker across all evaluated benchmarks. These results suggest that iterative evidence acquisition can benefit from operating over a structured multi-granularity representation, rather than repeatedly accessing visual evidence from the original temporal sequence.

Comparison with Hierarchical Video Representation. VideoTree provides a particularly relevant comparison because it also organizes long-video content hierarchically. Under the same Qwen2.5-VL-7B backbone and 128-frame visual budget, HieraSeek outperforms VideoTree by 6.8 points on LongVideoBench, 3.9 points on Video-MME overall, and 3.1 points on LVBench. These gains suggest that hierarchical organization alone is not sufficient; coupling the visual representation with learned multi-round search enables more effective access to task-relevant evidence during reasoning.

4.3 Ablation Studies

Representation Structure and Policy Learning. We disentangle the contributions of representation structure and policy learning by comparing uniformly sampled frames with the proposed multi-granularity hierarchy under both SFT and SFT+RL. As shown in Table 2, the hierarchy consistently improves over the flat representation under SFT across all six benchmarks, demonstrating the effectiveness of structured multi-granularity representation. RL benefits both variants, with the strongest performance obtained when policy optimization is combined with the hierarchy. The larger gap between the hierarchical and flat variants after RL suggests that the two components are complementary: the hierarchy provides a structured search space, while the learned policy exploits it for evidence localization.

Evidence-Seeking Reward. We isolate the contribution of the evidence-seeking objective on LongVideoBench using the same SFT initialization and hierarchical representation. As shown in Table 3, removing either R_{seek} or the trajectory factor $c(T)$ degrades performance. The larger drop without R_{seek} confirms the value of explicitly supervising intermediate evidence localization, while $c(T)$ helps discourage unnecessarily long search trajectories.

Table 3: Ablation of the evidence-seeking reward on LongVideoBench.

Method	LongVideoBench
SFT	56.6
SFT+RL w/o R_{seek}	57.4
SFT+RL w/o $c(T)$	58.0
SFT+RL (Full)	**59.2**

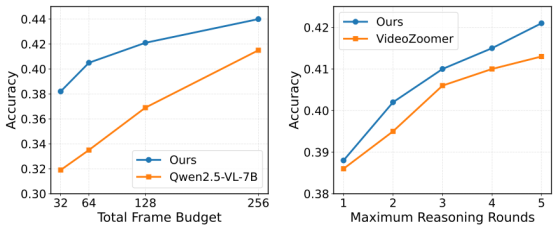


Figure 4: Test-time analysis on LVBench.

Table 4: End-to-end inference efficiency on VideoMME under the same 128-frame visual budget. Latency is averaged per sample, and wall-clock time is measured over the full evaluation.

Method	Output Tokens	FLOPs (T)	Latency (s)	Peak Memory (GB)	Wall-Clock (h)
Qwen2.5-VL	67.58	1319.41	24.23	37.32	2.55
VideoZoomer	412.57	1430.66	35.74	42.53	3.36
HieraSeek w/o reuse	207.26	1402.76	32.89	40.39	3.09
HieraSeek	**207.26**	**1383.62**	**29.69**	45.51	**2.79**

Training Dynamics. Fig. 3 shows that the evidence-seeking reward and answer accuracy improve over RL training, while the average number of search rounds decreases. This trend suggests that the policy learns to localize evidence more effectively rather than simply increasing the amount of exploration.

4.4 Efficiency Analysis

We evaluate end-to-end inference efficiency on VideoMME under the same 128-frame visual budget. In addition to Qwen2.5-VL and VideoZoomer, we include a variant of HieraSeek without intermediate-state reuse to isolate the cost of repeated visual encoding. All measurements use 8 NVIDIA A100 GPUs. Table 4 shows that HieraSeek incurs additional cost over single-pass inference, but is substantially more efficient than VideoZoomer among multi-round methods, with fewer generated tokens and lower latency. Intermediate-state reuse further reduces inference time by avoiding repeated visual encoding across search rounds, at the cost of increased GPU memory for caching. This exposes a clear compute–memory trade-off in the reuse mechanism. A detailed latency breakdown is provided in Appendix F.

4.5 Test-Time Analysis

Fig. 4 examines HieraSeek under different test-time budgets on LVBench. Performance improves with both additional visual evidence and additional search rounds, while HieraSeek maintains an advantage over the corresponding baselines across the tested settings. The gains gradually saturate as the budgets increase, indicating that the learned policy can benefit from additional test-time resources without requiring excessive exploration.

5 Conclusion

We presented HieraSeek, a long-video reasoning framework that turns the visual representation itself into a searchable structure. By constructing a multi-granularity hierarchy within the ViT encoder and learning to access it through multi-round evidence search, HieraSeek integrates visual representation and evidence acquisition within a unified reasoning process. Across six long-video understanding and reasoning benchmarks, the proposed framework consistently improves over strong single-round and multi-round baselines, while intermediate-state reuse reduces repeated visual computation during search. These results support searchable visual representations as a promising direction for scalable and efficient reasoning over long videos.

AI USE STATEMENT

We used generative AI tools for limited language polishing and editorial assistance in preparing the manuscript. We also used Gemini 2.5 Pro to synthesize multi-round reasoning trajectories for constructing the supervised fine-tuning dataset, as described in the paper. These generated trajectories were automatically validated and filtered based on action validity, consistency with annotated evidence intervals, and answer correctness. The research ideas, methodology, experiments, analyses, and conclusions were developed and verified by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed implementation and training settings in the main paper and appendix. In the supplementary material, we will release the full code pipeline covering data generation and preprocessing, supervised fine-tuning, reinforcement learning, and evaluation. We will also provide executable training and inference scripts, configuration files, and a README with step-by-step instructions for reproducing the reported experiments.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

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

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. *arXiv preprint arXiv:2408.10188*, 2024b.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. *arXiv preprint arXiv:2512.22315*, 2025.

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

Chaoyou Fu, Yuhan Dai, 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*, pp. 24108–24118, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

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. In *International Conference on Learning Representations*, volume 2026, pp. 74904–74933, 2026.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Computing Surveys*, 57(10):1–35, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

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.

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*, pp. 5971–5984, 2024.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2794–2804, 2023.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Yuanbin Man, Ying Huang, Chengming Zhang, Bingzhe Li, Wei Niu, and Miao Yin. Adacmˆ 2: On understanding extremely long-term video with adaptive cross-modality memory reduction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8534–8544. IEEE, 2025.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. *arXiv preprint arXiv:2510.20470*, 2025.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. Longvideo-r1: Smart navigation for low-cost long video understanding. *arXiv preprint arXiv:2602.20913*, 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

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. *arXiv preprint arXiv:2410.17434*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6099–6113, 2025.

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*, pp. 22958–22967, 2025a.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025b.

Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10467–10475, 2026.

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*, pp. 3272–3283, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

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

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing "thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*, 2025.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Xinlei Yin, Xiulian Peng, Xiao Li, Zhiwei Xiong, and Yan Lu. Hierarchical long video understanding with audiovisual entity cohesion and agentic search. *arXiv preprint arXiv:2601.13719*, 2026a.

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*, pp. 24042–24051, 2026b.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024a.

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*, 2024b.

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*, pp. 22056–22065, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Mmr-v: What's left unsaid? a benchmark for multimodal deep reasoning in videos. In *International Conference on Learning Representations*, volume 2026, pp. 67949–67991, 2026.

A Analysis of Temporal Redundancy and Initial Sampling

We provide additional analyses of temporal redundancy across ViT depths and the effect of the initial sampling density. These analyses motivate the progressive representation construction introduced in Sec. 3.2 and the use of up to 4,096 initially sampled frames.

Inter-Frame Similarity across ViT Depths. We use the visual encoder of Qwen2.5-VL and uniformly sample 128 frames from each video in the long-video subset of Video-MME. Following Sec. 3.2, we represent each frame at the ℓ-th ViT block using the spatially averaged attention key vectors, denoted as $F_t^{(\ell)}$. For a temporal distance d, we measure the average cosine similarity between pairs of frame representations separated by d:

$$\mathrm{Sim}^{(\ell)}(d) = \frac{1}{N-d} \sum_{t=1}^{N-d} \cos\left(F_t^{(\ell)}, F_{t+d}^{(\ell)}\right). \tag{10}$$

As shown in Fig. 5(a), temporally nearby frames consistently exhibit higher similarity than frames separated by larger temporal distances, indicating substantial local redundancy. The similarity profiles also vary across ViT depths, motivating progressive representative-frame selection across multiple encoder blocks rather than a single compression operation at a fixed representation level.

Effect of Initial Sampling Density. We further examine how the initial number of uniformly sampled frames affects temporal redundancy before hierarchy construction. As shown in Fig. 5(b), increasing the sampling density gradually leads to diminishing changes in the measured frame similarity, with the trend becoming relatively stable around 4,096 sampled frames. We therefore use up to 4,096 initial frames as a practical balance between temporal coverage and the computation required for hierarchy construction.

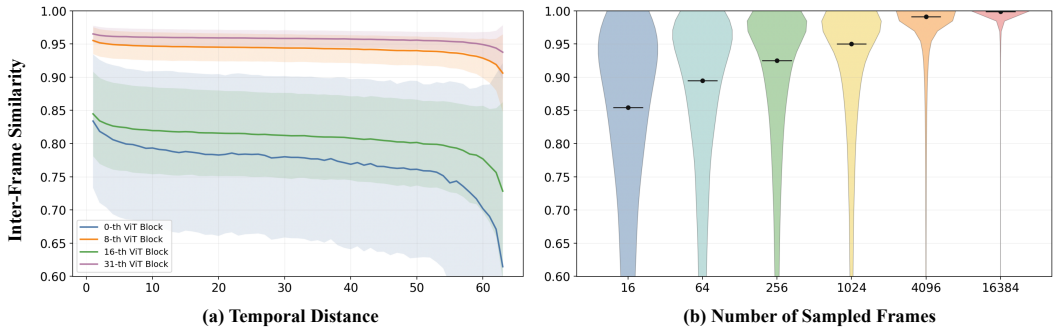


Figure 5: Analysis of temporal redundancy and initial sampling density. **Left:** cosine similarity between frame-level key representations as a function of temporal distance at representative ViT depths. **Right:** effect of the number of initially sampled frames on inter-frame similarity.

B Greedy Approximation for Key-Guided Medoid Selection

The subset-selection objective in Eq. 4 requires searching over all subsets of k frames and is therefore expensive to solve exactly. We use a greedy approximation that incrementally selects representative frames according to their reduction of the objective.

For a temporal group $g_{\ell,j}$, we first compute the pairwise distances $D_j^{(\ell)}(a,b)$ defined in Eq. 3. We initialize the representative set with the frame that minimizes the total distance to all frames in the group:

$$b^* = \arg\min_{b\in\{1,\dots,n\}} \sum_{a=1}^{n} D_j^{(\ell)}(a,b). \tag{11}$$

We then maintain, for each frame a, its distance to the nearest selected representative,

$$d_a = \min_{b\in S_{\ell,j}} D_j^{(\ell)}(a,b). \tag{12}$$

Algorithm 1 Greedy Key-Guided Medoid Selection

Require: Temporal group $g_{\ell,j} = \{f_1, \ldots, f_n\}$, representative budget k
Ensure: Representative index set $S_{\ell,j}$

Compute frame representations $\{F_{j,a}^{(\ell)}\}_{a=1}^{n}$
Compute pairwise distances $D_j^{(\ell)}(a, b)$ using Eq. 3
$b^* \leftarrow \arg\min_b \sum_{a=1}^{n} D_j^{(\ell)}(a, b)$
$S_{\ell,j} \leftarrow \{b^*\}$
for $a = 1, \ldots, n$ **do**
  $d_a \leftarrow D_j^{(\ell)}(a, b^*)$
end for
while $|S_{\ell,j}| < k$ **do**
  for all $u \notin S_{\ell,j}$ **do**
    $\Delta(u) \leftarrow \sum_{a=1}^{n} \left[d_a - \min(d_a, D_j^{(\ell)}(a, u)) \right]$
  end for
  $u^* \leftarrow \arg\max_{u \notin S_{\ell,j}} \Delta(u)$
  $S_{\ell,j} \leftarrow S_{\ell,j} \cup \{u^*\}$
  for $a = 1, \ldots, n$ **do**
    $d_a \leftarrow \min(d_a, D_j^{(\ell)}(a, u^*))$
  end for
end while
return $S_{\ell,j}$

At each subsequent step, for every unselected candidate u, we compute the reduction in the selection objective that would result from adding u:

$$\Delta(u) = \sum_{a=1}^{n} \left[d_a - \min\left(d_a, D_j^{(\ell)}(a, u) \right) \right]. \tag{13}$$

The candidate with the largest reduction is added to the representative set, and the nearest-representative distances are updated accordingly. This process is repeated until k representatives have been selected. Algorithm 1 summarizes the procedure.

C Temporal Interval Mapping

Given a temporal interval predicted by the search policy, we map it to a compact set of nodes in the visual hierarchy. Let the queried interval be $I = [l, r]$, which is first converted to the corresponding frame-index range $[i_l, i_r]$ over the initially sampled video.

Because the interval boundaries may not align with the temporal groups used during progressive representation construction, we allow boundary expansion with a maximum padding budget P. The padding budget limits the number of additional frames included outside the originally requested interval.

We first collect the lowest-level groups whose temporal support covers the query interval. We then traverse the hierarchy upward. Whenever two covered neighboring nodes are the left and right children of the same parent, they are replaced by their parent node. If further merging is prevented only by boundary misalignment, neighboring groups may additionally be included as long as the total number of padded frames does not exceed P, allowing the corresponding sibling nodes to be merged at a higher level. This process continues until no further valid merging is possible.

The resulting node set,

$$\mathcal{C} = \mathrm{Cover}_{\mathcal{H}}(I; P),$$

therefore consists of one or more highest-level hierarchy nodes whose combined temporal support covers the requested interval, subject to the padding budget. These nodes are then used to retrieve the cached intermediate visual states for the current search round.

D Evidence-Seeking Reward Details

We provide the action-specific definitions used by the evidence-seeking reward in Sec. 3.4. Let $I_t = [l_t, r_t]$ denote the temporal interval at search round t, $I_{t+1} = [l_{t+1}, r_{t+1}]$ the interval reached after action a_t, and $I^* = [l^*, r^*]$ the annotated evidence interval. The evidence coverage ratio is

$$\rho_t^{\text{cov}} = \frac{|I_{t+1} \cap I^*|}{|I^*|}. \tag{14}$$

`zoom_in`. A `zoom_in` action is valid when the new interval is a strict sub-interval of the current search region and the current region contains the annotated evidence:

$$I_{t+1} \subsetneq I_t, \qquad I^* \subseteq I_t. \tag{15}$$

The reduction ratio is measured relative to the current interval:

$$\rho_t^{\text{red}} = \frac{|I_{t+1}|}{|I_t|}. \tag{16}$$

Thus, among actions that sufficiently cover the evidence, a smaller resulting interval receives a larger search reward.

`shift_to`. The `shift_to` action is used to relocate the search rather than further refine the current interval. To evaluate the amount of search-space reduction after relocation, we identify the most recent interval I_x in the search history that contains both the new interval and the annotated evidence:

$$I_{t+1} \subseteq I_x, \qquad I^* \subseteq I_x. \tag{17}$$

We then define

$$\rho_t^{\text{red}} = \frac{|I_{t+1}|}{|I_x|}. \tag{18}$$

Using I_x as the reference avoids measuring a relocation against an unrelated current interval and instead evaluates how much the search is refined relative to their most recent shared temporal context.

For both actions, the step reward is

$$q_t = \mathbb{1}[\text{valid}(a_t)]\, \mathbb{1}[\rho_t^{\text{cov}} \geq \eta]\, \rho_t^{\text{cov}} \left(1 - \rho_t^{\text{red}}\right), \tag{19}$$

where η is the minimum evidence-coverage threshold. The evidence-seeking reward is applied only to intermediate search actions (`zoom_in` and `shift_to`); the final `answer` action is evaluated separately by the task-level accuracy reward R_{acc}.

E Additional Training Details

E.1 SFT Data Construction

We construct supervised fine-tuning data from the *long-video reflection* subset of VideoSIAH (Yang et al., 2025) and CG-Bench (Chen et al., 2024a). For each question, we use Gemini 2.5 Pro (Comanici et al., 2025) to generate a multi-round search trajectory following the action space defined in Sec. 3.3.

At each reasoning round, the teacher is provided with visual frames sampled from the current temporal interval, their approximate timestamps, and the reasoning history from previous rounds. The number of frames provided to the teacher is adjusted according to the duration of the current interval, with a maximum of 256 frames:

$$N_{\text{frame}} = \min\left(256, \begin{cases} \max(12, \lceil 3T \rceil), & T \leq 8, \\ \max(24, \lceil 2T \rceil), & 8 < T \leq 30, \\ \max(32, \lceil 1.25T \rceil), & 30 < T \leq 200, \\ 256, & T > 200, \end{cases}\right), \tag{20}$$

where T denotes the duration of the current interval in seconds.

Both datasets provide temporal evidence annotations. To encourage diverse search trajectories, we vary the round at which the annotated evidence interval is exposed to the teacher as private guidance. Specifically, samples are divided into three groups with a ratio of $30\%/40\%/30\%$, for which the evidence hint becomes available from rounds 1, 2, and 3, respectively.

The system prompt used for teacher-trace generation is shown in Fig. 9. It instructs the teacher model to reason only from the currently visible video segment and previous trajectory, produce grounded visual observations, and output either a tool call or a final answer following a strict tag-based format.

We automatically validate the generated trajectories at each round. We check the output format and action validity, verify that the predicted temporal regions are consistent with the annotated evidence, and require the final answer to match the ground-truth answer. We additionally remove trajectories with highly repetitive reasoning across adjacent rounds. After filtering, the resulting SFT dataset contains approximately 10K multi-round trajectories.

E.2 Training Data Distribution

Fig. 6 summarizes the round-wise distribution of the constructed SFT trajectories and action types. Most trajectories terminate within the first few reasoning rounds, while later rounds contain progressively fewer samples. The action distribution reflects the multi-round search process, with `zoom_in` and `shift_to` used for evidence localization and `answer` terminating the trajectory.

E.3 SFT and Implementation Details

We perform SFT on Qwen2.5-VL-7B-Instruct using LLaMA-Factory (Zheng et al., 2024). We train for 3 epochs with a learning rate of 1×10^{-5} using a cosine learning-rate schedule and a warmup ratio of 0.03. The per-device batch size is 1, with a global batch size of 8 across 8 NVIDIA A100 GPUs. The maximum sequence length is 23,552 tokens.

Initial Sampling and Hierarchy Construction. For each video, we uniformly sample up to 4,096 frames with a maximum resolution of 224×224 per frame. These frames enter the early ViT blocks for hierarchy construction, where progressive frame selection and temporal merging reduce the visual sequence across encoder depth. We apply progressive pruning over the first six ViT blocks. The temporal regularization weight in Eq. 3 is set to $\alpha_{\text{temp}} = 0.1$. At each pruning stage, each temporal group retains half of its frames before the retained frames from neighboring groups are merged for the next stage.

The 4,096-frame input therefore refers to the initial visual information used to construct the hierarchy and does not imply that all frames are processed through the full vision encoder. As pruning proceeds, progressively fewer visual tokens are forwarded through later ViT blocks.

Intermediate-State Caching. For each hierarchy node, we cache the intermediate visual states produced at the ViT block where that node is constructed. When a node is retrieved during multi-round search, its cached states are forwarded through the subsequent ViT blocks rather than recomputing the corresponding frames from the beginning of the visual encoder. States associated with hierarchy nodes are retained for reuse throughout the search trajectory, while local visual tokens from previous reasoning rounds are not explicitly kept in the language-model context.

Visual-Budget Accounting. The initial hierarchy-construction budget and the reasoning-time visual-context budget refer to different stages of the pipeline. During evaluation, the cumulative visual context provided to the language model across search rounds is limited to 128 pre-temporal-merge frame equivalents. Qwen2.5-VL applies a temporal merging factor of two, so this corresponds to a temporal dimension equivalent to 64 frames at the LLM input. This reasoning-time budget should not be interpreted as the total number of raw frames processed by the visual encoder, since HieraSeek may initially sample up to 4,096 frames for hierarchy construction. The associated encoder-side computation is accounted for separately in our end-to-end FLOPs, latency, and memory measurements in Sec. 4.4.

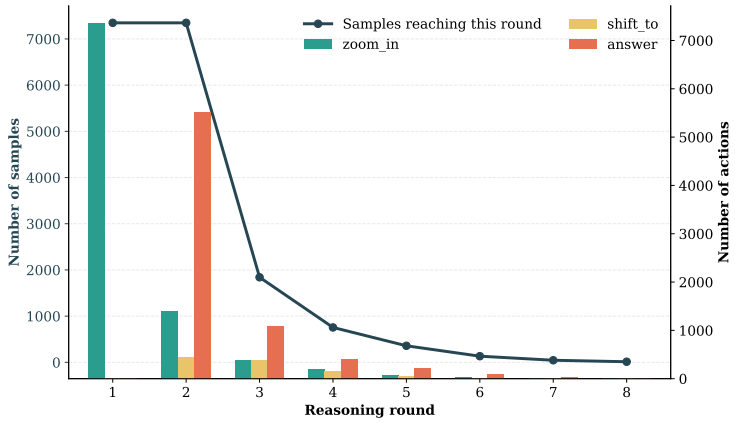


Figure 6: Round-wise distribution of training samples and action types in the constructed multi-round reasoning data. The top panel shows the number of samples reaching each reasoning round, and the bottom panel shows the number of actions of each type.

Table 5: Latency breakdown of HieraSeek with and without intermediate-state reuse. All values are averaged per sample and reported in seconds.

Method	**Build/Cache**	**Medoid Sel.**	**Visual Re-encode**	**Total**
HieraSeek w/o reuse	2.6	0.3	4.3	32.89
HieraSeek	2.6	0.3	**1.1**	**29.69**

E.4 RL Training

We further optimize the SFT model using Dr. GRPO (Liu et al., 2025) with the verl framework (Sheng et al., 2025). The RL data are drawn from CG-Bench (Chen et al., 2024a) and LongVideoReason (Chen et al., 2025).

Training is conducted on 8 NVIDIA A100 GPUs. For each optimization step, we use a global prompt batch size of 12 and sample 16 rollouts per prompt. Multi-round generation is limited to at most $T_{\max} = 8$ search rounds. For response generation, we use vLLM (Kwon et al., 2023) with a sampling temperature of 1.0. The actor learning rate is 1×10^{-6}, and the entropy coefficient is set to 0.001.

The RL objective combines the task-level accuracy reward with the evidence-seeking reward introduced in Sec. 3.4. Importantly, R_{seek} is applied only to intermediate `zoom_in` and `shift_to` actions, while the final `answer` action is evaluated solely by R_{acc}.

E.5 Prompt Template

We use a unified action format for both training and inference. At each round, the model is instructed to either issue a `zoom_in` or `shift_to` action with a target temporal interval, or produce the final answer with `answer`. The complete prompt templates used for teacher-trajectory generation and model training are provided in Fig. 8.

F Efficiency Breakdown

To better understand the source of the efficiency gains from intermediate-state reuse, we decompose the inference latency of HieraSeek into three components: initial hierarchy construction and caching, greedy medoid selection, and additional visual re-encoding during multi-round search. All measurements follow the same VideoMME evaluation setting used in Sec. 4.4.

The hierarchy construction and medoid-selection costs are identical for the two variants and account for only a small portion of the overall inference time. The main benefit of intermediate-state reuse

comes from reducing additional visual re-encoding during search, from 4.3 s to 1.1 s per sample. This confirms that the latency reduction observed in Table 4 primarily results from avoiding repeated visual encoding rather than from changes to the hierarchy construction procedure.

G CASE STUDY

We further present several representative case studies from LVBench to illustrate the reasoning behaviors learned by HieraSeek under different long-video reasoning scenarios. As shown in Fig. 7, 10, and 11, the videos and questions are sampled from LVBench, and HieraSeek can flexibly combine global understanding, local refinement, and temporal relocation during multi-round reasoning.

Fig. 7 shows a direct-answer example where the model identifies globally visible textual evidence ("London 2012") and correctly answers the question without additional navigation. This demonstrates that the model can avoid unnecessary exploration when sufficient evidence already exists in the global context.

Fig. 10 illustrates a `zoom_in`-based reasoning trajectory. Starting from the entire video, the model progressively narrows the temporal interval around the sandwich-eating scene to inspect the subsequent actions in finer detail, eventually identifying that the vlogger enjoys the sun outside the house. This example highlights the effectiveness of hierarchical local refinement for evidence localization.

Fig. 11 presents a more challenging example requiring temporal relocation. The model initially explores an incorrect interval, then recognizes the mismatch between the current evidence and the question, and finally invokes `shift_to` to move to another temporal region containing the relevant character introduction. By recovering from the mislocalized search trajectory, the model successfully identifies Sierra's former roommate. This example demonstrates the ability of HieraSeek to perform corrective exploration and robust multi-step reasoning over long videos.

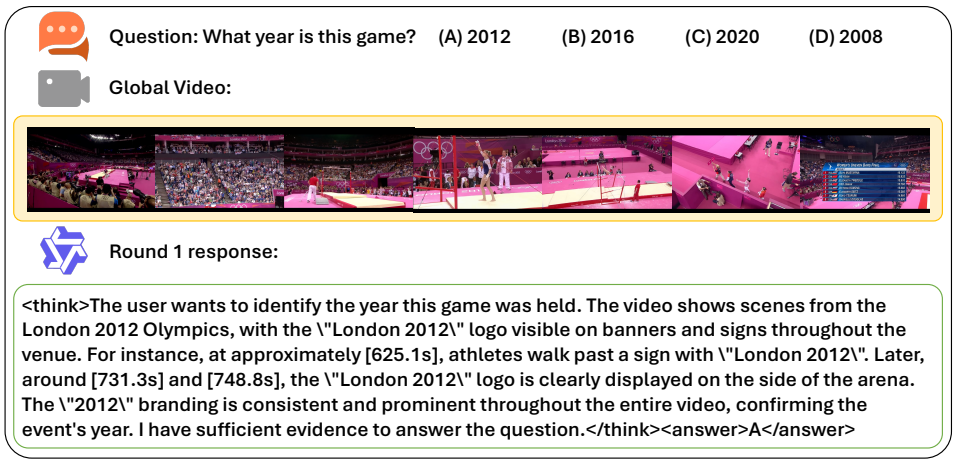


Figure 7: Case study of direct answering. Given the globally visible evidence, HieraSeek directly identifies the "London 2012" visual cues in the video and answers the question without invoking additional navigation actions.

Unified Prompt Template for Training and Inference

```
<|im_start|>system
You are a helpful assistant.

# Tools
You may call one or more functions to assist with the user query.
You are provided with function signatures within <tools></tools>
XML tags:

<tools>
{"type": "function", "function": {"name": "zoom_in", ...}}
   Inspect a narrower sub-interval inside the current video
interval.

{"type": "function", "function": {"name": "shift_to", ...}}
   Move to another video interval when the current interval is not
the best place to continue.
</tools>

For each function call, return a JSON object with function name and
arguments within <tool_call></tool_call> XML tags:
<tool_call>{"name": <function-name>, "arguments":
<args-json-object>}</tool_call>

Write your reasoning inside <think>...</think>; if more evidence
is needed, output one <tool_call>...</tool_call>, otherwise give a
concise final answer inside <answer>...</answer>.
<|im_end|>

<|im_start|>user
{Question}
<video><|im_end|>

<|im_start|>assistant
```

Figure 8: Unified prompt template used for both training and inference. The prompt specifies the available navigation tools and constrains the model to produce structured reasoning, tool calls, or final answers in a fixed format.

System Prompt for Teacher-Trace Generation

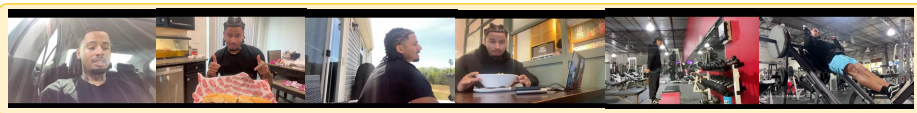
```
You are an expert long-video analyst creating high-quality teacher
traces for tool-augmented video QA distillation.
You only have access to:
1.  the already-finished transcript from earlier rounds, and
2.  the frames sampled from the CURRENT interval of the video for
this round.
Base every statement on visible evidence from those frames or on
observations already written in earlier rounds.  Treat earlier
transcript as memory, not content to paraphrase repeatedly.  If
you already paraphrased the question or goal in earlier rounds,
do not restate it every round.  Usually paraphrase once in round
1; mention it again only when it materially helps this round's
understanding, reasoning, or action choice.
When choosing a tool call, make sure <think> provides a coherent
bridge to that action:  give grounded visual evidence supporting
the destination interval; if grounded evidence is insufficient,
state an explicit uncertainty-driven hypothesis for why that
destination is still promising.  Each round should prioritize
newly grounded observations, updated uncertainty, and next-step
decisions.
Never mention hidden controller guidance, reference annotations,
dataset metadata, answer keys, evidence timestamps, or that the
interval policy is supervised.  Never disclose tool-budget metadata
in <think>, including shift_to total count, used count, remaining
count, quota, or similar controller-oriented phrasing.  Never say
that you are looking at frames, images, prompts, screenshots, or
annotations.  Describe the content as a continuous silent video
segment.
Available actions:
<tools>
{"name":  "zoom_in", "args":  ["start_time", "end_time"]}
   Inspect a stricter sub-interval inside the current interval.

{"name":  "shift_to", "args":  ["start_time", "end_time"]}
   Move to a different interval when the current focus appears
mislocalized.
</tools>
Strict output rules for every round:
- Output exactly one <think>...</think> block first.
- After </think>, output exactly one more block:  either
<tool_call>...</tool_call> or <answer>...</answer>.
- Do not output any other tags, markdown fences, headings, bullet
lists, or prose outside those tags.
- The <think> block must contain 3 to 8 sentences of concrete
visual reasoning.
- If you use <tool_call>, the JSON must be compact and valid.
- If you use <answer>, it must contain only the final answer, with
no explanation.
- Never claim to have seen the destination interval of a tool call
before the next round actually shows it.
```

Figure 9: System prompt used for generating SFT teacher traces. The prompt constrains the teacher model to produce grounded visual reasoning and strictly formatted tool-augmented trajectories.

Question: What does the vlogger do after finish eating the sandwiches?
(A) Get work done in a coffee shop (B) Work out in a gym
(C) Enjoy the sun outside the house (D) Talk to the audience in the car

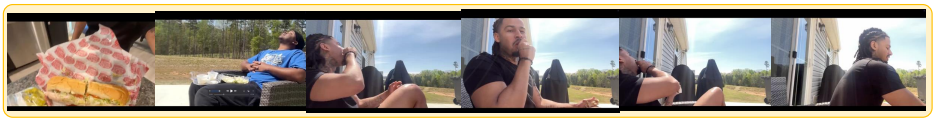
Global Video:



Round 1 response:

<think>The user wants to know what the vlogger does after eating sandwiches. I need to locate the scene where he eats sandwiches and then observe his subsequent actions. The current interval is very broad, covering the entire video. I can see the vlogger in various settings: in a car, at a mall, at a restaurant, outside on a porch, in a coffee shop, and at a gym. Around [626.6s], he is sitting on a porch eating. Shortly after, around [640.9s], he is seen outside on a sunny day, sitting on a chair next to the house. This appears to be the event described in the question. To confirm this and see the transition more clearly, I will zoom in on the interval surrounding this event. I'll select a window that starts before he gets the food and ends after he's done eating and enjoying the weather. A focus on the interval from 600 to 800 seconds should provide the necessary detail.</think><tool_call>{\"name\":\"zoom_in\",\"arguments\":{\"start_sec\":600.0,\"end_sec\":800.0}}</tool_call>

Round 2 Video:



Round 2 response:

<think>This interval provides a clear sequence of events. The vlogger is seen eating sandwiches at a table inside a house from approximately [600.2s] to [616.7s]. He then leaves the house, as seen at [617.8s], and is next shown sitting on a lounge chair outside on a porch at [623.3s]. From [628.8s] onwards, he is outside on a sunny day, sitting on a chair next to the house and enjoying the weather. He is also seen drinking from a cup and rubbing his face with a towel. This directly answers the question. The evidence is sufficient to select the correct option.</think><answer>C</answer>

Figure 10: Case study of `zoom_in`-based local refinement. HieraSeek first observes the full video, then zooms into the interval around the sandwich-eating scene to inspect the subsequent events, correctly identifying that the vlogger enjoys the sun outside the house.

Question: What is the name of Sierra's former roommate?
(A) Mia (B) Kindall (C) Alexis (D) Roxana

Global Video:



Round 1 response:

<think>The user is asking for the name of Sierra's former roommate. I need to locate the part of the video where this is discussed. I can see a segment titled \"Real Girlfriend vs. Online Girlfriend\" starting around [500.9s]. Within this segment, at approximately [959.6s], a lower-third graphic appears that reads \"Says Brian promised to marry her\". This \"her\" is likely Sierra's roommate. The name \"Brian\" is also seen on a sign at [1456.6s]. To confirm the name of the roommate, I need to find the introduction of this character. The characters are introduced with lower-third graphics at various points. I will zoom in on the section of the video that covers this entire \"Real Girlfriend vs. Online Girlfriend\" segment to find the specific introduction of Sierra's roommate. A good interval to investigate would be from the beginning of this segment to its conclusion. I'll select a range that comfortably covers all the potential introductions and the related storylines.</think><tool_call>{\"name\":\"zoom_in\",\"arguments\":{\"start_sec\":500.0,\"end_sec\":1500.0}}</tool_call>

Round 2 Video:



Round 2 response:

<think>I have now examined the interval from [500.0s] to [1500.0s]. This segment, titled \"Real Girlfriend vs. Online Girlfriend,\" features several characters. I've confirmed the name \"Brian\" from a sign at [1456.6s]. The question asks for the name of Sierra's former roommate. At [959.6s], a lower-third graphic appears stating \"Says Brian promised to marry her,\" with \"her\" presumably being Sierra. However, this segment also introduces other roommates, including one named \"Katie\" around [1013.4s]. The question is ambiguous. It's possible the roommate's name is mentioned in a different part of the video. I recall seeing another segment about \"Tutu Many Girlfriends\" earlier in the video, around [208.0s]. I will shift my focus to that earlier part of the video to see if the roommate's name is mentioned there, as it might provide context not present in the \"Real Girlfriend vs. Online Girlfriend\" segment. I'll investigate the interval from 200 to 400 seconds.</think><tool_call>{\"name\":\"shift_to\",\"arguments\":{\"start_sec\":200.0,\"end_sec\":400.0}}</tool_call>

Round 3 Video:



Round 3 response:

<think>I have now examined the interval from [200.0s] to [400.0s]. This segment, titled \"Tutu Many Girlfriends,\" introduces several characters. At [356.3s], a man is identified by a lower-third graphic as \"DAVID\". He is accused of cheating on a woman named \"Sierra\" with another woman, \"Kindall\", who is also a roommate. This provides a direct link between David and Kindall. The question asks for the name of Sierra's former roommate. The on-screen text refers to Kindall as Sierra's roommate. However, the question is phrased ambiguously, asking for the name of the roommate Sierra *was* with. This could imply a past relationship. Given the visual evidence, David is the person accused of the affair, and he is shown with Sierra. It is highly probable that the question is referring to the name of the person David was involved with, which is Kindall. I have sufficient evidence to conclude the answer.</think><answer>B</answer>

Figure 11: Case study of `shift_to`-based temporal relocation. After initially focusing on an irrelevant interval, HieraSeek shifts to an earlier segment containing the key character introduction and correctly identifies Sierra's former roommate as Kindall.