TIMEQUEST: SCALING INFERENCE-TIME COMPUTE FOR VIDEO UNDERSTANDING VIA DECOUPLED TEMPORAL EVIDENCE ACQUISITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling inference-time compute has improved the performance of Large Language Models (LLMs), yet its application to video understanding is limited by a fundamental perception bottleneck. Existing Video-LLMs primarily scale reasoning in the text space; however, if a model misses brief but critical visual evidence during its initial pass, no amount of subsequent text-based "thinking" can recover it. In this paper, we introduce TIMEQUEST, a framework that decouples temporal evidence acquisition from reasoning via a novel Navigator–Oracle architecture. The Navigator, trained using Group Relative Policy Optimization (GRPO) with a smooth distance-based temporal reward, iteratively identifies informative video segments. Rather than naively stacking raw high-dimensional visual tokens (which rapidly exhausts the context window), the Oracle translates these localized segments into dense textual descriptions. This accumulated context is appended to the initial query, enabling the Oracle to process the original video alongside an evidence-augmented prompt to deduce the final answer. This design fundamentally shifts inference-time compute from raw visual processing to active evidence distillation. Extensive evaluations on standard and temporally grounded Video QA benchmarks demonstrate consistent improvements of up to **+6.0** accuracy points, confirming that active evidence acquisition is a key factor in effective inference-time scaling for video understanding.

1 INTRODUCTION

In complex video understanding, the primary bottleneck is rarely logical reasoning, but acquiring precise, temporally sparse visual evidence (Zhao et al., 2025; Rasekh et al., 2025). In long and untrimmed videos, the key signals required to answer a query often occur in brief and scattered intervals, making reliable evidence extraction a fundamental challenge for Video-LLMs.

Recently, scaling inference-time compute has emerged as a powerful paradigm for overcoming reasoning bottlenecks in Large Language Models (LLMs) (DeepSeek-AI et al., 2025; Wei et al., 2022) with significant performance gains in complex logical and mathematical tasks (Guo et al., 2025; de Winter et al., 2024). In text-only settings, this scaling is achieved by autoregressively generating intermediate reasoning tokens before formulating a final answer. This paradigm has been extended to Video-LLMs, marking a transition from models optimized for direct instruction following (such as the Qwen2-VL series) (Wang et al., 2024; Bai et al., 2025b; Li et al., 2024) to models that explicitly generate reasoning traces (such as Qwen3-VL-Thinking and Qwen3.5) (Bai et al., 2025a; Qwen Team, 2026; Team et al., 2025). However, while reasoning tokens improve logical synthesis, directly porting this text-centric paradigm to multimodal tasks exposes a fundamental limitation.

The temporal nature of video data presents a unique challenge to the paradigm of inference-time compute scaling (Luo et al., 2025). While real-world and complex benchmarking videos frequently span several minutes, the decisive visual cues often occupy only brief, scattered multi-second windows. If a model fails to pay attention to these fleeting events during an initial pass, no amount of subsequent text-based deliberation can recover the lost visual signal; a model simply cannot deduce what it has not adequately observed. Consequently, prolonging the computation phase over an incomplete visual representation yields diminishing returns. Thus, for Video-LLMs, inference-time

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.

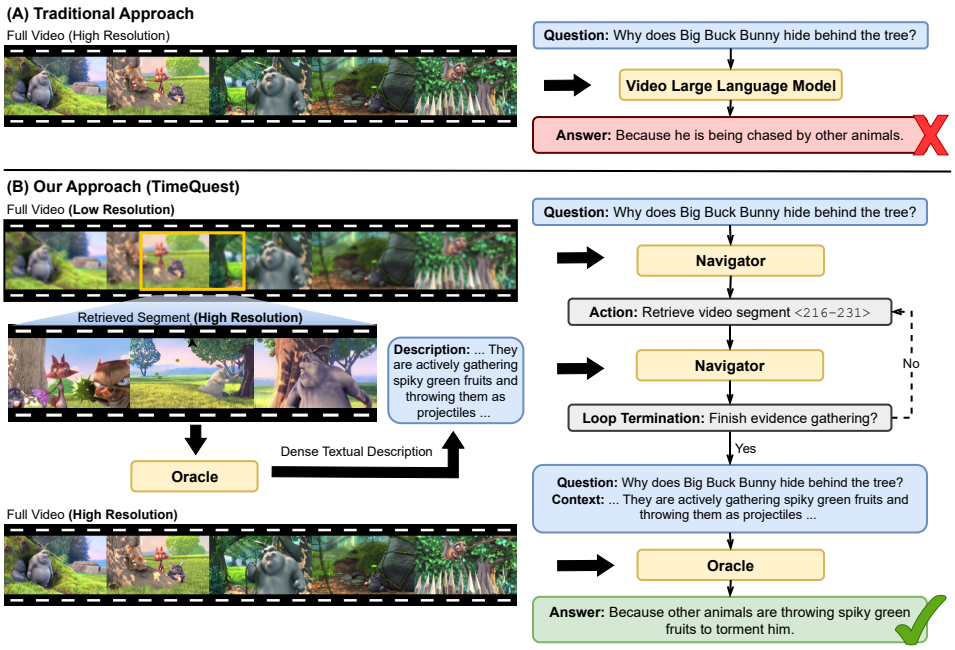


Figure 1: **(A) Traditional Approach:** Directly processing the full high-resolution video often misses brief, crucial evidence. **(B) TIMEQUEST:** A *Navigator* iteratively scans the low-resolution video to retrieve targeted temporal segments across multiple turns (one cycle in this example). An *Oracle* translates these brief high-resolution clips into dense textual descriptions. This accumulated evidence trace is appended to the original question, enabling the Oracle to accurately deduce the final answer from an evidence-augmented prompt.

compute must be fundamentally redefined: rather than exclusively generating text tokens, computational resources should be dynamically allocated toward active, targeted visual evidence gathering.

Recognizing the need for active visual exploration, a nascent body of concurrent work has attempted inference-time scaling for Video-LLMs by framing evidence gathering as a multi-turn process optimized via Reinforcement Learning (RL) (Li et al., 2025a; Wang et al., 2025d; Ding et al., 2026; He et al., 2026; Zhang et al., 2025). However, these methods are constrained by a critical architectural bottleneck: they entangle temporal navigation and logical deduction within a monolithic model. As these models iterate, they continually append raw, high-dimensional visual tokens to their context window. This rapid context inflation not only causes lost-in-the-middle phenomena (Liu et al., 2024), but severely restricts the maximum number of exploration turns a model can execute before hitting strict memory or context window limits.

To address these limitations, we propose TIMEQUEST[1], a framework that redefines multimodal inference-time scaling by decoupling visual evidence acquisition from logical reasoning via a novel Navigator–Oracle architecture (Figure 1). To avoid naively accumulating high-dimensional visual tokens (which rapidly exhausts the context window), a lightweight *Navigator* dynamically identifies informative temporal segments across multiple turns. A high-capacity *Oracle* then translates these localized segments into dense textual descriptions. Once exploration concludes, this accumulated evidence trace is appended to the initial question. The Oracle then processes the original video alongside this evidence-augmented prompt to formulate its final answer. This structural shift distills fleeting visual signals into a highly token-efficient context, equipping the Video-LLM with targeted information and enabling extended exploration trajectories.

To achieve autonomous exploration, the Navigator is optimized via Group Relative Policy Optimization (GRPO) using a smooth distance-based temporal reward. Unlike from prior interactive frameworks that rely on sparse outcome or Intersection over Union (IoU) rewards (Li et al., 2025b;

[1]We provide our code anonymously as Supplementary Material and will make it publicly available when accepted.

He et al., 2026; Shen et al., 2025), our continuous reward formulation enables more stable and effective training. Furthermore, TIMEQUEST is a highly modular enhancement natively compatible with any off-the-shelf Video-LLM. Extensive evaluations demonstrate that our framework yields consistent improvements of up to **+6.0** accuracy points across standard Video Question Answering and temporally grounded Video QA benchmarks.

2 RELATED WORK

2.1 INFERENCE-TIME SCALING IN VIDEO-LLMS

The primary challenge in video understanding lies in the trade-off between managing extensive visual contexts and retaining fine-grained temporal details. Early Video-LLMs relied on static strategies to handle this, employing frame sampling (Yu et al., 2023), explicit sequence compression (Qian et al., 2024; Zeng et al., 2025; Li et al., 2025c), or coarse-to-fine heuristics like *SlowFocus* (Nie et al., 2024). These methods operate within a rigid, single-pass computational graph where the visual signal is fixed before reasoning begins. More recently, scaling inference-time compute has revolutionized text-centric reasoning (Snell et al., 2026; Guo et al., 2025; OpenAI, 2024). Extending this paradigm to the multimodal domain, works such as Feng et al. (2025); Wang et al. (2025c); Li et al. (2025b); Chen et al. (2025) utilize Group Relative Policy Optimization (GRPO) to elicit inference-time deliberation. However, these RL frameworks remain largely single-turn, processing a fixed set of frames without dynamically acquiring new visual evidence.

2.2 DYNAMIC EVIDENCE ACQUISITION AND THE MONOLITHIC BOTTLENECK

Recognizing the limitations of static perception, a concurrent line of research has introduced multi-turn, interactive frameworks (He et al., 2026; Zhang et al., 2025; Ding et al., 2026; Liu et al., 2026; Xie et al., 2025; Shen et al., 2025; Li et al., 2026). Models such as *VITAL* and *VideoMind* empower Video-LLMs to actively search long videos by iteratively retrieving localized temporal segments. Despite their dynamic nature, these methods couple navigation and reasoning in a monolithic model, repeatedly appending raw visual tokens to the context. This context inflation limits searchable steps due to token and memory constraints. By contrast, TIMEQUEST introduces a fundamental structural decoupling. Instead of accumulating raw visual pixels, our lightweight Navigator translates targeted video segments into concise textual evidence. This scales the reasoning engine's "visual context" efficiently, allowing the frozen Oracle to deduce answers over a highly dense, unified textual space and bypass the visual context bottleneck entirely.

2.3 REWARD FORMULATION FOR TEMPORAL NAVIGATION

The effectiveness of RL in multi-turn evidence acquisition is heavily dependent on reward formulation. Existing frameworks targeting dynamic evidence acquisition (Li et al., 2025b; He et al., 2026; Li et al., 2025d; Ding et al., 2026; Shen et al., 2025), predominantly optimize their exploration policies using sparse, discrete rewards. These typically include terminal binary QA accuracy or strict Intersection over Union (IoU) thresholds. While sufficient for end-to-end task completion, sparse rewards struggle to provide the dense credit assignment required for a model to learn stable, sequential temporal navigation across long-form video. To overcome this, TIMEQUEST optimizes its Navigator via GRPO using a smooth distance-based temporal reward. By explicitly rewarding the continuous temporal proximity of the Navigator's actions to the ground-truth evidence, we achieve more stable training.

3 METHOD

TIMEQUEST redefines VideoQA inference-time scaling by shifting computation toward *active temporal evidence acquisition*. A lightweight Navigator proposes informative temporal segments, which a frozen high-capacity Oracle converts into concise textual evidence. The accumulated evidence is then used by the Oracle for final answering, avoiding repeated expansion of the raw visual context.

3.1 TASK DEFINITION

Given a video V, a natural-language question q, and a set of candidate answers A, the task of *VideoQA with temporal grounding* is to (1) select the correct answer $a \in A$ and (2) identify one or more temporal segments that serve as supporting *evidence*.

In TIMEQUEST, evidence collection is handled by the **Navigator**. It outputs a set of timestamp intervals $\mathcal{S} = \{(s_j, e_j)\}_{j=1}^{m}$ with $0 \leq s_j \leq e_j \leq T$, where T is the video duration in seconds. Each segment $V[s_j : e_j]$ should contain the event, action, interaction, object, or dialogue needed to answer q. In practice, the Navigator scans the full video using a low-resolution view to efficiently propose candidate evidence intervals, while high-resolution processing is reserved for short retrieved clips.

Since long videos include substantial irrelevant content, selecting the right moments is critical. TIMEQUEST produces predicted evidence intervals $\widehat{\mathcal{S}}$ and then passes the corresponding localized descriptions to a **frozen Oracle**, which selects the final answer $\hat{a} \in A$ using the accumulated evidence context.

3.2 MODEL INTERACTION AND RUNTIME LOOP

At each cycle $k \geq 1$, the loop follows a fixed tag-based protocol: (1) the Navigator outputs a segment request using `<timestamp>` tags, e.g., `<timestamp> <11.9> <19.9> </timestamp>`; (2) the system retrieves the corresponding subvideo $V_{t_k} \subset V$ (as a short high-resolution clip) and passes it to the frozen Oracle; (3) the Oracle returns a localized description wrapped in `<sentence>` tags, which the Navigator appends to its growing evidence context; and (4) the Navigator either proposes another `<timestamp>` or emits `<end>` to stop evidence collection.

Final answer. After `<end>`, the frozen Oracle receives the original video V and question q together with the accumulated sentences $\{s_1, \ldots, s_k\}$, and selects the final answer $\hat{a}$.

Figure 2 shows an illustration of the per-cycle generation process and a worked example of the reward calculations. See Appendix A for full prompt examples

3.3 TRAINING PIPELINE

We train the Navigator in two stages, SFT followed by GRPO. In both stages, LoRA adapters are inserted into its vision encoder and language backbone, while the Oracle remains frozen.

3.3.1 STAGE 1: SUPERVISED FINE-TUNING (SFT)

We first perform supervised fine-tuning on annotated multi-turn transcripts that follow our runtime protocol: `<timestamp>` $\rightarrow$ retrieve segment $\rightarrow$ Oracle returns `<sentence>` $\rightarrow \cdots \rightarrow$ `<end>`. This stage serves primarily as a *warm start*, teaching the Navigator the required interaction format and the overall evidence-collection task (i.e., how to issue valid timestamp requests, incorporate returned sentences, and terminate).

During SFT, timestamp requests and the `<end>` decision are teacher-forced, while the Oracle-provided sentences are included as observations in the transcript. This initializes the policy with the correct multi-turn structure before GRPO, ensuring stable rollouts and preventing early failures due to malformed tags or incorrect loop behavior.

3.3.2 STAGE 2: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Starting from the SFT checkpoint, we refine the Navigator with Group Relative Policy Optimization (GRPO). We view the Navigator as a sequential policy π that produces rollouts consisting of timestamp turns followed by a terminating `<end>`. Importantly, GRPO optimizes the Navigator *only* for evidence collection quality: rewards are computed from timestamp correctness and the termination decision, not from final answer accuracy. This isolates temporal navigation from downstream answer generation, reduces reward variance, and directly targets precise localization. We apply group-relative normalization to stabilize updates, and update only the LoRA parameters, keeping training efficient while improving temporal grounding behavior.

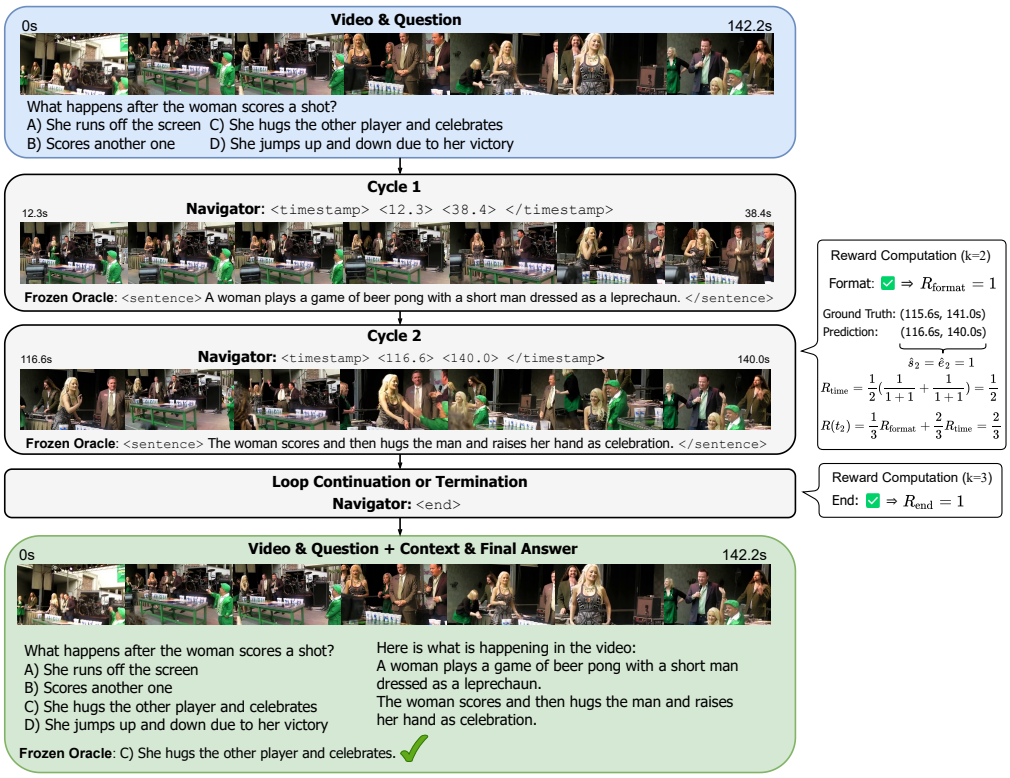


Figure 2: Multi-turn generation process and worked numeric examples of per-turn rewards. The image shows how timestamps and sentence turns evolve across cycles, terminating with an <end> tag before passing the context to the frozen base model for the final answer.

3.4 Reward Design

TimeQuest is designed to improve VideoQA by *collecting the right temporal evidence* for a frozen Oracle, rather than predicting the final answer itself. We therefore exclude answer-based rewards and optimize the Navigator purely for generating accurate timestamp queries and stopping at the correct time. The total episode return $R(\tau)$ is computed from per-turn rewards on timestamp proposals and a termination reward.

For each timestamp turn t_k, we use a weighted combination of a format reward and a temporal accuracy reward:

$$R(t_k) = 1/3 R_{\text{format}} + 2/3 R_{\text{time}}. \tag{1}$$

Here, R_{format} is a binary indicator that rewards valid <timestamp> syntax.

Temporal reward. Let $(\hat{s}_k, \hat{e}_k)$ be the predicted start/end times for the k-th timestamp, and (s_k^*, e_k^*) the ground-truth boundaries. We define boundary errors $\Delta s_k = |\hat{s}_k - s_k^*|$ and $\Delta e_k = |\hat{e}_k - e_k^*|$, and compute:

$$R_{\text{time}} = \frac{1}{2}\left(\frac{1}{1+\Delta s_k} + \frac{1}{1+\Delta e_k}\right). \tag{2}$$

Despite prior GRPO-based temporal supervision typically relying on IoU-style overlap rewards (Li et al., 2025b; He et al., 2026), we instead adopt this smooth, distance-based formulation. IoU assigns zero reward to non-overlapping proposals, which can yield sparse or vanishing learning signals early in training. In contrast, Eq. 2 provides a dense signal that reflects *how far* each boundary is from the target, enabling steady refinement toward the correct evidence interval.

Termination reward. We additionally reward the Navigator for ending evidence collection at the appropriate time. The *correct final step* is defined as the step immediately after all ground-truth

evidence timestamps in the training transcript have been issued. When the Navigator generates <end>, we assign:

$$R_{\text{end}} = \begin{cases} 1, & \text{if <end> is predicted at the correct final step} \\ 0, & \text{otherwise (e.g., premature or delayed termination).} \end{cases} \tag{3}$$

4 EXPERIMENTS

We evaluate whether scaling inference-time compute through *targeted visual evidence gathering* improves long-form VideoQA. Concretely, we compare a strong frozen Video-LLM used as the Oracle against the same Oracle augmented with TIMEQUEST, where a lightweight Navigator performs multi-turn temporal navigation and accumulates high-density textual evidence before providing the final answer. The experiments use 16 NVIDIA A100 GPUs.

4.1 EXPERIMENTAL SETUP

Backbones and Variants. We use Qwen2.5-VL 3B as the primary backbone for training the Navigator policy (SFT followed by GRPO). At evaluation time, TIMEQUEST is used as a plug-and-play temporal module that augments a frozen Oracle. We report results across multiple Oracle backbones (Qwen2.5-VL(Bai et al., 2025b), Qwen3-VL(Bai et al., 2025a), InternVL3(Zhu et al., 2025), InternVL3.5(Wang et al., 2025b), and LLaVA-OneVision 1.5(An et al., 2025)) in two variants: (i) Base, which answers directly from the video and question, and (ii) $\text{TIMEQUEST}_{\text{Base}}$, where the Navigator first gathers localized textual evidence through multi-turn timestamp proposals and the Oracle then produces the final answer from the original video alongside the accumulated evidence context.

Controlled Inference Budget. For each backbone, Base and $\text{TIMEQUEST}_{\text{Base}}$ are evaluated under the same model scale and decoding settings for the Oracle (e.g., identical max frames or pixels where applicable). The additional computation in TIMEQUEST comes from the Navigator's low-resolution scan of the full video and a small number of Oracle calls on short, retrieved *high-resolution* segments; importantly, we do not increase the Oracle's parameter count or alter its weights.

Training Data. We train the Navigator via the two-stage pipeline (SFT then GRPO) on the temporally annotated **Grounded-VideoLLM training dataset** Wang et al. (2025a), which contains samples derived from ActivityNet Captions(Krishna et al., 2017) and QVHighlights(Lei et al., 2021). The resulting training corpus contains 15,453 samples, which is relatively small compared to modern Video-LLM training datasets. Despite its modest size, this mixture provides diverse temporal patterns, including actions, events, and highlight-like moments, encouraging robust temporal localization and evidence retrieval.

Benchmarks. We evaluate on seven benchmarks spanning long-form multiple-choice VideoQA and grounding-focused VideoQA. We use EgoSchema (Mangalam et al., 2023), TVBench (Cores et al., 2025), LifeQA (Castro et al., 2020), CinePile (Rawal et al., 2024), Video-MME, and MLVU to measure long-form VideoQA over extended temporal contexts, and NExT-GQA (Xiao et al., 2024) to evaluate grounded VideoQA where both answer correctness and temporal evidence quality are assessed. Detailed results on Video-MME and MLVU are reported in Appendix B.6.

Evaluation Metrics. For EgoSchema, TVBench, LifeQA, CinePile, Video-MME, and MLVU, we report standard multiple-choice accuracy (%). For NExT-GQA, we also report grounding-aware metrics: Acc@QA measures final answer accuracy regardless of localization; Acc@GQA counts examples where the answer is correct *and* the predicted segment sufficiently matches the ground-truth evidence (IoP ≥ 0.5); IoP@τ (Intersection-over-Prediction) measures how much of the predicted segment is supported by the ground truth at threshold τ (precision-oriented); and IoU@τ (Intersection-over-Union) is the standard overlap metric between predicted and ground-truth segments at threshold τ.

Table 1: Accuracy (%) on long-form VideoQA benchmarks. 'TIMEQUEST$_{\text{backbone}}$' denotes the same backbone equipped as Oracle with our multi-turn evidence collection policy. * indicates results where EgoSchema scores are taken from (llm-stats, 2026), while TVBench and CinePile scores are reported from (Cores et al., 2025; Rawal et al., 2024), respectively. The TVBench score for GPT-4o is image-input only.

Model	EgoSchema	TVBench	LifeQA	CinePile
SlowFocus 7B(Nie et al., 2024)	39.7	–	–	–
Video-MTR 7B(Xie et al., 2025)	62.4	–	–	–
GPT-4o*	72.2	35.8	–	56.1
Gemini 2.0 Flash*	71.5	–	–	–
Gemini 1.5 Pro*	–	46.5	–	–
Qwen2.5-VL 7B	63.0	39.6	64.6	36.6
TIMEQUEST$_{\text{Qwen2.5-VL 7B}}$	**65.0**	**42.2**	**66.5**	**42.5**
Qwen3-VL 8B	71.8	48.6	83.3	51.4
TIMEQUEST$_{\text{Qwen3-VL 8B}}$	**72.9**	**49.5**	**84.6**	**52.9**
InternVL 3 8B	73.8	44.2	80.0	46.4
TIMEQUEST$_{\text{InternVL 3 8B}}$	**75.6**	**46.8**	**82.4**	**48.3**
InternVL 3.5 8B	69.0	47.5	83.2	46.1
TIMEQUEST$_{\text{InternVL 3.5 8B}}$	**71.6**	**48.9**	**83.9**	**48.9**
LLaVA-OneVision 1.5 8B	46.6	34.1	68.8	39.2
TIMEQUEST$_{\text{LLaVA-OneVision 1.5 8B}}$	**52.6**	**35.7**	**72.4**	**41.9**

4.2 RESULTS

4.2.1 LONG-FORM VIDEOQA ACCURACY

As shown in Table 1, TIMEQUEST improves long-form VideoQA accuracy across all evaluated Oracle backbones and benchmarks. The largest gains appear on CinePile, where adding TIMEQUEST raises Qwen2.5-VL from 36.6 to 42.5 and LLaVA-OneVision 1.5 from 39.2 to 41.9, indicating that targeted retrieval can be particularly useful when relevant information is distributed across long videos. On EgoSchema, TIMEQUEST improves LLaVA-OneVision 1.5 from 46.6 to 52.6 and Qwen3-VL from 71.8 to 72.9. Similar improvements are observed on TVBench and LifeQA across the stronger Oracle backbones.

These results show that the gains do not depend on modifying or enlarging the Oracle. Instead, the same frozen Video-LLM benefits from receiving localized evidence selected by the Navigator. This supports temporal evidence acquisition as an independent mechanism for improving long-form VideoQA.

4.2.2 GROUNDING QUALITY

To directly evaluate whether the system retrieves the *right* moments, we report grounding-aware metrics on NExT-GQA in Table 2. Across Oracles, TIMEQUEST improves temporal grounding and, in most cases, final answer accuracy by using the Navigator to retrieve evidence-bearing segments.

For Qwen2.5-VL, Acc@GQA increases from 41.0 to 47.0 and Acc@QA from 73.9 to 74.8, alongside large overlap gains such as IoP@0.3 from 61.0 to 85.2 and IoU@0.5 from 23.3 to 60.4. For Qwen3-VL, Acc@GQA rises from 31.2 to 44.7 while IoU@0.5 improves from 30.8 to 61.3. LLaVA-OneVision 1.5 shows a similar pattern, with Acc@GQA improving from 29.5 to 42.4 and IoU@0.5 increasing from 10.0 to 58.9.

For InternVL backbones, grounding metrics are not available for the Base variants because they do not reliably output timestamp segments under our prompting protocol. With TIMEQUEST, however, grounded segments are produced by the external Navigator, and answer accuracy improves from 76.8 to 80.1 for InternVL 3 and from 75.7 to 76.6 for InternVL 3.5. Overall, these results indicate that the temporal module substantially improves the quality of localized evidence supplied to the Oracle.

Table 2: Performance on NExT-GQA. 'TIMEQUEST$_{\text{backbone}}$' denotes the same backbone equipped as Oracle with our Navigator-based evidence collection. For InternVL backbones, Base grounding results are not reported because they do not reliably output timestamp segments under our prompting protocol.

Model	Acc@QA	Acc@GQA	IoP@0.3	IoP@0.5	IoU@0.3	IoU@0.5
Grounded-VideoLLM 4B(Wang et al., 2025a)	–	26.7	–	34.4	–	18.0
Zoom-Zero 7B(Shen et al., 2025)	–	29.0	–	–	55.6	33.8
TVG-R1 7B(Chen et al., 2025)	–	22.1	–	–	41.6	20.8
VideoChat-R1 7B(Li et al., 2025b)	70.6	24.3	–	–	50.2	27.7
VITAL 7B(Zhang et al., 2025)	78.7	–	–	–	–	–
Qwen2.5-VL 7B	73.9	41.0	61.0	52.4	41.7	23.3
TIMEQUEST$_{\text{Qwen2.5-VL 7B}}$	**74.8**	**47.0**	**85.2**	**63.8**	**84.7**	**60.4**
Qwen3-VL 8B	74.2	31.2	52.1	39.4	48.2	30.8
TIMEQUEST$_{\text{Qwen3-VL 8B}}$	**75.0**	**44.7**	**86.1**	**64.7**	**85.4**	**61.3**
InternVL 3 8B	76.8	–	–	–	–	–
TIMEQUEST$_{\text{InternVL 3 8B}}$	**80.1**	48.5	82.2	60.2	81.4	56.4
InternVL 3.5 8B	75.7	–	–	–	–	–
TIMEQUEST$_{\text{InternVL 3.5 8B}}$	**76.6**	48.8	84.2	63.2	83.6	59.7
LLaVA-OneVision 1.5 8B	66.6	29.5	46.9	41.8	19.6	10.0
TIMEQUEST$_{\text{LLaVA-OneVision 1.5 8B}}$	**67.2**	**42.4**	**83.6**	**62.2**	**82.8**	**58.9**

4.3 Ablation Studies and Analyses

We conduct a series of ablations and analyses to examine the key design choices and properties of TIMEQUEST. We first compare temporal evidence acquisition with text-space inference scaling, then study the effect of the temporal reward design and the computational efficiency of our decoupled framework. To further analyze the design choices and behavior of TIMEQUEST, we conduct a broader set of experiments in the appendix. These include ablations on the number of retrieval cycles, the choice between frame indices and continuous timestamps, and the visual budget allocated to retrieved segments; comparisons of prompted, SFT-only, and SFT+GRPO Navigators; analyses of how the retrieval budget adapts to different task types; a manual analysis of failure cases and the limitations of the textual evidence bottleneck; sensitivity to the maximum retrieval budget; evaluation on additional long-video benchmarks; and transfer experiments with substantially larger Oracle models. Together, these experiments provide a more detailed analysis of how TIMEQUEST collects evidence, allocates inference-time computation, and generalizes across tasks, training configurations, benchmarks, and Oracle scales. We present these analyses in Appendix B.

4.3.1 Evidence Acquisition vs. Text-Space Scaling

To compare temporal evidence acquisition with text-space inference scaling, we evaluate Qwen3-VL 8B Base and Thinking, with and without TIMEQUEST. As shown in Table 3, Thinking provides only modest gains over Base, whereas TIMEQUEST produces larger improvements, particularly in temporal grounding. Combining both achieves the strongest overall performance, suggesting that temporal evidence acquisition and text-space scaling are complementary.

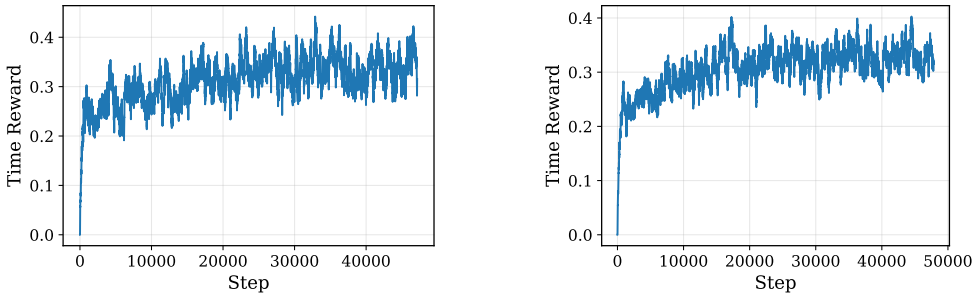
Table 3: Comparison of inference-time scaling strategies using Qwen3-VL 8B.

Model	EgoSchema	TVBench	NExT-GQA Acc@GQA	NExT-GQA IoU@0.5
Qwen3-VL 8B Base	71.8	48.6	31.2	30.8
Qwen3-VL 8B Thinking	72.0	48.8	35.1	35.3
TIMEQUEST$_{\text{Qwen3-VL 8B}}$	72.9	49.5	44.7	61.3
TIMEQUEST$_{\text{Qwen3-VL 8B Thinking}}$	**73.4**	**49.8**	**44.9**	**61.4**

4.3.2 Temporal Reward Design — Distance-Based vs. IoU

We compare our distance-based GRPO reward with an IoU-based reward, which depends on segment overlap and becomes zero for non-overlapping proposals.

As shown in Figure 3, the distance-based reward provides a smoother training signal early in training. Although both rewards reach about 0.4 by the end of epoch 1, their values are not directly comparable: IoU ≈ 0.4 indicates partial overlap, whereas a distance reward of ≈ 0.4 corresponds to boundary errors of roughly 1.5s.



(a) IoU reward (epoch 1).

(b) Distance-based reward (epoch 1).

Figure 3: Temporal reward training curves for the first epoch.

Table 4 shows that the distance-based reward improves accuracy from 77.32 to 81.21 (+3.89), despite a slightly lower mean reward (0.37 vs. 0.41). A mean distance reward of 0.37 corresponds to average boundary errors of roughly 1.7s. Unlike IoU, the distance-based reward therefore provides a non-zero localization signal even for proposals without overlap with the ground-truth interval.

Table 4: Temporal reward ablation on our internal test dataset.

Reward	**Accuracy (%)**	**Mean Reward**
IoU-based reward	77.32	0.41
Distance-based reward (ours)	81.21	0.37

4.3.3 Computational Efficiency

We analyze the theoretical FLOPs of a standard three-turn evidence-gathering trajectory. Unlike a monolithic architecture that repeatedly expands the high-resolution visual context, TimeQuest uses a lightweight Navigator and processes retrieved segments independently with the Oracle.

The monolithic 7B baseline requires 6,656 TFLOPs during prefill, compared with 1,935 TFLOPs for our 3B Navigator + 7B Oracle, a $3.4\times$ reduction. We assume Qwen2.5-VL 3B/7B, a 200-frame high-resolution video, a 100-frame low-resolution proxy, three 90-frame retrieval cycles, and standard patch tokenization (784 pixels/token). This highlights the computational benefit of converting retrieved visual content into compact textual evidence rather than repeatedly expanding the visual context.

5 Conclusion

We introduced TimeQuest, a framework that scales inference-time compute for VideoQA by shifting computation from long-context visual processing to *active temporal evidence gathering*. TimeQuest combines a lightweight Navigator that proposes segments in a multi-turn loop with a frozen Oracle that converts retrieved clips into compact textual evidence and answers from the accumulated context, avoiding the growth of high-dimensional visual tokens. The Navigator is trained with SFT+GRPO using a smooth distance-based temporal reward for reliable localization. Across benchmarks, TimeQuest consistently improves performance, including CinePile (Qwen2.5-VL: 36.6→42.5), EgoSchema (LLaVA-OneVision 1.5: 46.6→52.6), and grounded QA on NExT-GQA (Qwen2.5-VL Acc@GQA: 41.0→47.0; IoU@0.5: 23.3→60.4). A current limitation is the timestamp-only interaction protocol. Future work will enable richer evidence queries and more flexible evidence aggregation while keeping the Oracle frozen.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing portions of the manuscript, including improving grammar, spelling, word choice, clarity, concision, and organization. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, assist with proofs, propose or refine research hypotheses, design the methodology or experiments, implement the reported methods, translate research materials, clean or reformat datasets, perform qualitative or thematic data analysis, or interpret experimental results. The remaining required disclosure tasks are not applicable to this work.

We reviewed all AI-assisted content and independently verified the reported experimental results, numerical values, methodological descriptions, claims, and citations against the underlying work. The authors take full responsibility for the final content of this paper, including all text, claims, results, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe the proposed method and experimental setup in Sections 3 and 4, with additional ablations, implementation details, and experimental analyses provided in the appendix. The supplementary material includes the source code and experiment configurations used for training and evaluation. Our experiments use publicly available open-source models and datasets, and the reported settings are documented in the paper and accompanying code to facilitate reproduction of our results. The code will be made publicly available upon publication.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, et al. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. In *arXiv*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, et al. LifeQA: A Real-life Dataset for Video Question Answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4352–4358, 2020.

Ruizhe Chen, Zhiting Fan, Tianze Luo, Heqing Zou, Zhaopeng Feng, et al. Datasets and Recipes for Video Temporal Grounding via Reinforcement Learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 983–992, 2025.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in Time: A New Temporal Benchmark for VideoLLMs. In *Proceddings of the 36th British Machine Vision Conference*, 2025.

Joost C. F. de Winter, Dimitra Dodou, and Yke Bauke Eisma. System 2 Thinking in OpenAI's o1-Preview Model: Near-Perfect Performance on a Mathematics Exam. *Computers*, 13(11):278, 2024.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2025.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. VideoZoomer: Reinforcement-Learned Temporal Focusing for Long Video Reasoning. In *Proceedings of the International Conference on Learning Representations*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, et al. Video-R1: Reinforcing Video Reasoning in MLLMs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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 *Proceedings of the International Conference on Learning Representations*, 2026.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos, 2017. URL https://arxiv.org/abs/1705.00754.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting Moments and Highlights in Videos via Natural Language Queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, et al. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*, 2024.

Chenglin Li, Qianglong Chen, fengtao, and Yin Zhang. Iterative Zoom-In: Temporal Interval Exploration for Long Video Understanding. *arXiv preprint arXiv:2507.02946*, 2025a.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, et al. VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning. *arXiv preprint arXiv:2504.06958*, 2025b.

Yulin Li, Haokun Gui, Ziyang Fan, Junjie Wang, Bin Kang, et al. Less Is More, but Where? Dynamic Token Compression via LLM-Guided Keyframe Prior. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, et al. TempSamp-R1: Effective Temporal Sampling with Reinforcement Fine-Tuning for Video LLMs. *Advances in Neural Information Processing Systems (NeurIPS*, 2025d.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=FS3FzdrFZ7.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. VideoMind: A Chain-of-LoRA Agent for Temporal-Grounded Video Reasoning. In *Proceedings of the International Conference on Learning Representations*, 2026.

llm-stats. EgoSchema Benchmark Leaderboard . https://llm-stats.com/benchmarks/egoschema, 2026. Accessed: 2026-05-07.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When Thinking Drifts: Evidential Grounding for Robust Video Reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Ming Nie, Dan Ding, Chunwei Wang, Yuanfan Guo, Jianhua Han, et al. SlowFocus: Enhancing Fine-grained Temporal Understanding in Video LLM. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

OpenAI. OpenAI o1 System Card, 2024. URL https://openai.com/index/openai-o1-system-card/.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, et al. Streaming Long Video Understanding with Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 119336–119360, 2024.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, 2026. URL https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-05-07.

Ali Rasekh, Erfan Bagheri Soula, Omid Daliran, Simon Gottschalk, and Mohsen Fayyaz. Enhancing Temporal Understanding in Video-LLMs through Stacked Temporal Attention in Vision Encoders. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. CinePile: A Long Video Question Answering Dataset and Benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-Zero: Reinforced Coarse-to-Fine Video Understanding via Temporal Zoom-in. *arXiv preprint arXiv:2512.14273*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. In *Proceedings of the International Conference on Learning Representations*, 2026.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. *arXiv preprint arXiv:2507.01006*, 2025.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, et al. Grounded-VideoLLM: Sharpening Fine-grained Temporal Grounding in Video Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, et al. Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-RTS: Rethinking Reinforcement Learning and Test-Time Scaling for Efficient and Enhanced Video Reasoning. *arXiv preprint arXiv:2507.06485*, 2025d.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, and brian andothers ichter. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24824–24837, 2022.

Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. Can I Trust Your Answer? Visually Grounded Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-MTR: Reinforced Multi-Turn Reasoning for Long Video Understanding, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-Chained Image-Language Model for Video Localization and Question Answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, et al. TimeSuite: Improving MLLMs for Long Video Understanding via Grounded Tuning. In *Proceedings of the International Conference on Learning Representations*, pp. 38057–38081, 2025.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, et al. Thinking With Videos: Multimodal Tool-Augmented Reinforcement Learning for Long Video Reasoning. *arXiv preprint arXiv:2508.04416*, 2025.

Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, et al. Needle In A Video Haystack: A Scalable Synthetic Evaluator for Video MLLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

A PROMPT EXAMPLES

This section provides representative prompts used in TIMEQUEST for (i) localized evidence extraction by the frozen Oracle and (ii) final answer generation conditioned on the accumulated textual evidence. In both cases, the Oracle is prompted in natural language and remains frozen; the Navigator controls which temporal segment is provided for evidence extraction and when to terminate the evidence-gathering loop.

Box 1: Oracle prompt for localized `<sentence>` generation

You are an oracle to a Navigator. The Navigator has been asked to answer this question about a video with options:

Question: A person opens a refrigerator and then quickly closes it. What happens next?
(A) They start cooking on the stove.
(B) They take out a drink and pour it into a glass.
(C) They leave the room and turn off the lights.
(D) They sit down and read a book.

Now you have been given part of the video to explain for the Navigator.
Explain this video.

Box 2: Oracle prompt for final answer generation with accumulated evidence

Question: A person opens a refrigerator and then quickly closes it. What happens next?
(A) They start cooking on the stove.
(B) They take out a drink and pour it into a glass.
(C) They leave the room and turn off the lights.
(D) They sit down and read a book.

Here are some important events that happened in the video. You can use them to answer the question:

- The person opens the refrigerator, looks inside briefly, and closes it.
- They grab a bottle from the counter, take a glass, and pour liquid into it.

Answer the question with the option's letter from the given choices.

Box 1 shows the Oracle prompt used during the multi-turn evidence collection loop. The Navigator proposes a temporal segment, the system retrieves that subclip, and the frozen Oracle generates a localized textual description to be appended to the growing evidence context.

Box 2 shows the final answering prompt. After the Navigator emits `<end>`, the accumulated `<sentence>` descriptions (tool context) are provided to the Oracle alongside the original multiple-choice question, and the Oracle outputs the predicted option letter.

B Additional Ablation Studies

Unless stated otherwise, all ablations in this section use **Qwen2.5-VL 3B** as the frozen *Oracle*. We further analyze temporal representation, visual budget, evidence collection, Navigator training, task-dependent adaptiveness, failure modes, retrieval budget, broader benchmark performance, and Oracle scaling.

B.1 Grounding Units — Frame Indices vs. Seconds

TimeQuest operates by proposing temporal intervals via `<timestamp>` tags. We ablate the representation of these proposals: predicting discrete *frame indices* (e.g., `<i> <j>`) versus continuous *seconds*. We run SFT+GRPO with different GRPO step budgets and evaluate on EgoSchema (Table 5).

Table 5: Frame-index vs. second-based temporal proposals on EgoSchema across GRPO steps. Values are accuracy (%).

Setup	Frames (%)	Seconds (%)
SFT+GRPO (10k steps)	55.4	58.0
SFT+GRPO (25k steps)	56.4	57.0
SFT+GRPO (35k steps)	54.2	56.8

Second-based proposals outperform frame indices at all three training budgets, with the largest gap at 10k steps (58.0 vs. 55.4). The result is consistent with the use of continuous temporal intervals in the annotations and reduces sensitivity to video sampling density.

B.2 Constant vs. Variable Visual Budgets

We ablate the number of frames provided for each retrieved segment. We compare (i) a **constant-frame** setting that always samples a fixed number of frames regardless of segment duration, against (ii) a **constant-FPS (variable-frame)** setting where the number of sampled frames scales with segment length up to a maximum. Both settings use the same **SFT+GRPO** recipe and are evaluated after 5,000 GRPO steps.

Table 6: Constant vs. FPS-based (variable) visual budget on EgoSchema (SFT+GRPO, 5,000 steps). `const` uses a fixed number of frames per segment (120 frames average). `fps/max` samples at a fixed FPS so the number of frames varies with segment length (90 frames average observed).

Setup	EgoSchema Accuracy (%)
const (120 frames avg)	50.2
fps/max (90 frames avg)	50.4

The variable-frame setting gives a small accuracy improvement (50.4 vs. 50.2) while using fewer frames on average (90 vs. 120), suggesting a slightly more favorable compute–accuracy trade-off under this configuration.

B.3 Segment Relevance — TimeQuest-Guided vs. Fully Random Segments

We next test whether the gains come from providing the Oracle with more video or with more relevant evidence. Both settings use the same number and duration of segments; only the selection strategy differs: TimeQuest-guided timestamps are conditioned on the question and accumulated evidence, whereas the random baseline samples segments uniformly from the timeline.

TimeQuest-guided retrieval improves accuracy from 76.10 to 81.21 (+5.11), indicating that additional visual context alone is not sufficient: the retrieved segments must also be relevant to the question. The Navigator thus contributes through both temporal localization and evidence selection.

Table 7: Segment relevance ablation on our internal test dataset using Qwen2.5-VL 3B as the frozen Oracle. Both settings use the same number and duration of retrieved segments.

Retrieved segments	**Accuracy (%)**
Fully random segments	76.10
TIMEQUEST-guided segments (ours)	81.21

B.4 NAVIGATOR TRAINING: PROMPTED VS. SFT+GRPO

To isolate the contribution of the learned navigation policy, we replace the trained Navigator with the same pretrained Qwen2.5-VL 3B model used only through prompting. Both variants operate under the same retrieval protocol. As shown in Table 8, the trained Navigator achieves higher EgoSchema accuracy while using slightly fewer retrieval turns on average.

Table 8: Prompted vs. trained Navigator on EgoSchema. The prompted baseline uses the pretrained Qwen2.5-VL 3B model without SFT or GRPO.

Navigator	**EgoSchema (%)**	**Avg. retrieval turns**
Prompted Qwen2.5-VL 3B	60.2	2.3
Trained TIMEQUEST Navigator	**61.6**	**2.2**

The improvement suggests that the gains of TIMEQUEST are not explained solely by adding a pretrained 3B model to the inference pipeline. Learning the temporal-navigation policy through the proposed SFT+GRPO procedure provides an additional benefit while requiring slightly fewer retrieval turns.

B.5 TRAINING-STAGE ABLATION

We isolate the contribution of the two training stages by comparing the frozen Oracle alone, the Oracle augmented with an SFT-only Navigator, and the full SFT+GRPO Navigator. Results are reported on EgoSchema and the internal test set used for this analysis.

Table 9: Ablation of the two Navigator training stages.

Setup	**EgoSchema (%)**	**Internal Test (%)**
Oracle only	54.20	74.44
+ TIMEQUEST (SFT only)	53.60	76.21
+ TIMEQUEST (SFT+GRPO)	**58.00**	**81.16**

SFT provides the warm start needed to learn the interaction protocol, but SFT alone does not consistently improve all evaluation settings. Adding GRPO yields the largest improvement, increasing EgoSchema from 53.60 to 58.00 and the internal test accuracy from 76.21 to 81.16 relative to SFT alone. This is consistent with the role of SFT as format and behavior initialization and GRPO as the stage that directly optimizes temporal evidence acquisition.

B.6 BROADER BENCHMARK EVALUATION: VIDEO-MME AND MLVU

To evaluate whether the observed gains generalize beyond our main benchmark suite, we additionally evaluate Qwen2.5-VL 7B and TIMEQUEST on Video-MME without subtitles and both the Dev and Test splits of MLVU. We also include recently proposed inference-time temporal methods where results are available under the corresponding evaluation settings. As shown in Tables 10 and 11, TIMEQUEST improves the Qwen2.5-VL 7B Oracle by 3.3 points on Video-MME, 3.7 points on MLVU Dev, and 14.4 points on MLVU Test.

On Video-MME, TIMEQUEST improves over the Qwen2.5-VL 7B Oracle by 3.3 points and over Zoom-Zero by 2.4 points. It also exceeds the other listed methods under the reported evaluation

Table 10: Results on Video-MME without subtitles.

Model	Video-MME
VideoMindLiu et al. (2026)	58.2
Video-MTRXie et al. (2025)	62.7
Zoom-ZeroShen et al. (2025)	66.0
VITALZhang et al. (2025)	64.1
VideoZoomerDing et al. (2026)	65.2
Qwen2.5-VL 7BBai et al. (2025b)	65.1
Qwen2.5-VL 7B + TIMEQUEST	**68.4**

setting, with the exception that comparisons across methods should be interpreted in light of their respective reported protocols.

Table 11: Results on the Dev and Test splits of MLVU. 'NR' denotes that a result was not reported.

Model	Dev	Test
Video-MTRXie et al. (2025)	NR	50.4
Zoom-ZeroShen et al. (2025)	70.8	NR
VideoZoomerDing et al. (2026)	NR	55.8
Qwen2.5-VL 7B	70.2	45.5
Qwen2.5-VL 7B + TIMEQUEST	**73.9**	**59.9**

On MLVU, TIMEQUEST improves the Qwen2.5-VL 7B Oracle by 3.7 points on the Dev split and 14.4 points on the Test split. These results further support that targeted temporal evidence acquisition provides gains beyond the benchmarks used in the main evaluation.

B.7 TRANSFER TO A LARGER ORACLE

To evaluate whether the Navigator transfers across Oracle scales, we pair the same 3B Navigator with a frozen Qwen3-VL-32B-Instruct Oracle. As shown in Table 12, the combined system achieves 79.9% on Video-MME without subtitles, compared with 76.6% for the published Qwen3-VL-32B-Instruct result.

Table 12: Transfer of the 3B Navigator to a substantially larger frozen Oracle on Video-MME without subtitles.

Model	Video-MME (%)
Qwen3-VL-32B-Instruct	76.6
Qwen3-VL-32B-Instruct + TIMEQUEST	**79.9**

The 3B Navigator improves the substantially larger 32B Oracle by 3.3 points, indicating that its benefit is not limited to the 7B–8B Oracle scale used in our main experiments. Since the Navigator is considerably smaller than the Oracle, the improvement also suggests that the gains do not simply arise from adding model capacity; instead, the additional computation is used to acquire and provide targeted temporal evidence to the frozen Oracle. This transfer demonstrates that a lightweight Navigator can remain useful even when paired with a much larger Video-LLM.

B.8 SENSITIVITY TO THE MAXIMUM RETRIEVAL BUDGET

We further evaluate how performance changes with the maximum number of retrieval turns available to the Navigator. Because the Navigator may emit `<end>` before reaching the maximum budget, we report both the maximum allowed retrieval turns and the realized average computation.

Increasing the maximum retrieval budget from one to three turns gives incremental gains, while the adaptive policy achieves the highest accuracy with an average of 2.2 realized retrieval turns. This indicates that the policy does not simply benefit from always making the maximum number of

Table 13: Sensitivity to the maximum retrieval budget on EgoSchema. Avg. realized turns' counts Navigator retrieval turns, while Avg. Oracle calls' additionally includes the final Oracle call used for answer generation.

Maximum retrieval turns	Avg. realized turns	Avg. Oracle calls	EgoSchema (%)
1	1.0	2.0	49.1
2	1.7	2.7	49.6
3	1.9	2.9	49.7
Adaptive	2.2	3.2	**50.4**

retrieval requests; instead, allowing the Navigator to terminate adaptively provides a useful trade-off between additional evidence and computation.

B.9 Task-Dependent Adaptiveness

To characterize how TimeQuest allocates computation across different temporal tasks, we measure the average number of retrieval cycles on the ten TVBench task categories. As shown in Table 14, tasks requiring evidence from multiple temporal locations generally trigger more retrieval cycles. For example, Action Sequence and Action Count average 3.4 and 3.2 cycles, respectively, whereas Moving Direction and Action Antonym require only 1.5 and 1.2 cycles. This indicates that the Navigator adaptively allocates more retrieval effort to tasks with more distributed temporal evidence rather than applying a fixed budget to every input.

Table 14: Average number of retrieval cycles across TVBench task categories.

TVBench task	Avg. retrieval cycles
Action Sequence	3.4
Action Count	3.2
Egocentric Sequence	3.0
Object Shuffle	2.8
Unexpected Action	2.7
Object Count	2.6
Scene Transition	2.3
Action Localization	2.1
Moving Direction	1.5
Action Antonym	1.2

B.10 Failure Analysis

The textual evidence bottleneck can discard fine-grained visual information. This is a deliberate trade-off: TimeQuest replaces repeated accumulation of high-dimensional visual tokens with compact textual evidence, while the final Oracle still receives the complete original video together with the accumulated descriptions (Section 3). Thus, the descriptions summarize retrieved evidence without removing access to the original visual input.

To characterize the remaining limitations, we manually analyzed 50 failure cases from the three TVBench categories where Qwen3-VL 8B + TimeQuest underperforms the Base model. As shown in Table 15, most failures arise in Egocentric Sequence and Unexpected Action, with smaller degradation on Scene Transition.

Table 15: TVBench categories where Qwen3-VL 8B + TimeQuest underperforms the Base model.

Task	Base	TimeQuest	Difference
Egocentric Sequence	40.5	26.0	-14.5
Unexpected Action	36.6	20.7	-15.9
Scene Transition	86.5	84.3	-2.2

We assign each failure to one primary source, as shown in Table 16. The largest group (46%) involves information lost during textual conversion, primarily exact counts, fine-grained object attributes, and subtle spatial relations. Another 32% results from incomplete or incorrect temporal retrieval, while 22% occurs despite correct retrieval and description, indicating errors in the final Oracle answer generation.

Table 16: Primary failure sources among 50 manually analyzed failure cases.

Failure source	**Cases**	**Share**
Information lost during textual conversion	23	46%
Incomplete or incorrect temporal retrieval	16	32%
Correct evidence but incorrect final reasoning	11	22%

A representative failure occurs for questions requiring exact counts. For example, the retrieved clip may contain the necessary evidence, but the Oracle summarizes it as "The player dribbles the ball several times before taking a shot," losing the exact count required by the question. In contrast, temporal semantics such as event ordering and action localization are generally better preserved. These findings suggest that the main limitation of the textual bottleneck is preserving precise low-level visual details, rather than temporal evidence itself.

Despite these category-level failures, TIMEQUEST improves overall TVBench accuracy from 48.6% to 49.5%, indicating that gains on other task categories compensate for these weaknesses.