

MSTree: Multi-Scale Evidence Tree for Training-Free Long-Form Video Question Answering

Chen Dongxian^{1*}, Mingshuai Liu^{1*}, Yijie Peng¹, Hengrui Liu¹, Xin Li^{1,2†}

¹School of Information and Cyber Security,

²Public Security Big Data Strategy Research Center,

People's Public Security University of China, Beijing, China

{2024211497, liumingshuai, 2024211512, 202221440031}@stu.ppsuc.edu.cn

lixin@ppsuc.edu.cn

Abstract

Long-form video question answering hinges on locating brief, sparse evidence buried in minutes of dense action. Flattening the entire video into a single caption stream forces the LLM to sift through thousands of tokens, where decisive cues are easily drowned out by surrounding context. We present MSTree, a training-free framework that organizes captions into a Multi-Scale Evidence Tree. The tree is built by a coordinated dual-modal segmentation method that combines visual scene transitions and textual coherence under temporal contiguity, producing segments that are both visually and semantically consistent. Inference follows a mine-then-reason workflow: the LLM first scans coarse segment summaries to locate the most relevant event, then reasons over its fine-grained captions. When the first-round prediction falls below a confidence threshold, Contrastive Targeted Reflection (CTR) activates. The model articulates a discriminative gap that specifies what evidence is still missing to separate the top candidate answers, and an answer-anchoring constraint requires any newly retrieved evidence to both support the revised prediction and contradict the original. On EgoSchema, NExT-QA, and Video-MME, MSTree reaches 68.4%, 77.5%, and 55.7% accuracy, outperforming all training-free baselines.

1 Introduction

Video Question Answering (VideoQA) asks a system to integrate visual and textual content to answer natural language queries. Long-form video understanding is a particularly challenging setting: the difficulty arises less from raw video duration than from the need to locate brief, sparse evidence segments buried in minutes of dense action.

Trained video-language models (Cheng et al., 2024; Ranasinghe et al., 2025; Wang et al., 2024a;

Bai et al., 2025; Wang et al., 2025a; Li et al., 2024; Park et al., 2024) achieve strong performance but require heavy computation and struggle with the temporal context of very long videos. Training-free alternatives are a practical option. Image-grid approaches (Kim et al., 2024; Lin et al., 2023) uniformly sample frames for a vision-language model (VLM), but may miss brief critical moments and quickly exhaust the context window. The caption-then-reason paradigm (Zhang et al., 2024) sidesteps these visual bottlenecks by converting video into frame-level text and relying on a large language model (LLM) for long-range reasoning. However, supporting evidence in VideoQA is rarely distributed uniformly across a video; it typically concentrates in a small number of short, localized segments. Feeding an LLM a flat text stream spanning thousands of tokens therefore risks the “lost in the middle” effect (Liu et al., 2024), where sparse evidence is drowned out by contextual noise.

Hierarchical tree structures (Wang et al., 2025b) address this flat-caption bottleneck. By organizing captions at multiple levels of granularity, they enable a mine-then-reason workflow: the LLM first traverses coarse summaries to identify the relevant event, then examines fine-grained captions within it, reducing contextual noise (Figure 1).

The effectiveness of this workflow, however, depends on how the hierarchy is constructed. Existing hierarchical retrieval methods commonly construct their structures through K-means visual clustering (Wang et al., 2025b), enabling effective query-adaptive content organization. However, this construction does not explicitly optimize segment boundaries for both temporal contiguity and textual coherence. We therefore propose a coordinated dual-modal segmentation method in which visual anchors provide representative scene-level references, while a text-guided dynamic program refines the boundaries between them under temporal contiguity constraints, yielding segments that are

*Equal contribution.

†Corresponding author.

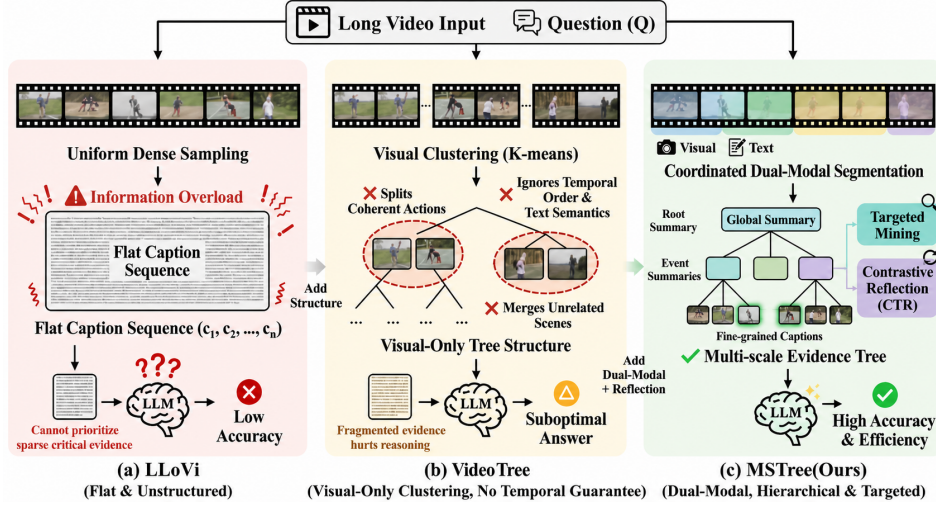


Figure 1: Motivation. Flat caption sequences expose all content equally, forcing the LLM to sift through irrelevant information to locate sparse evidence. MSTree organizes the same captions into a multi-scale hierarchy so that coarse-level summaries directly guide evidence mining, and Contrastive Targeted Reflection supplements missing evidence while suppressing noise.

both visually consistent and semantically cohesive.

Even a well-constructed evidence tree may not resolve subtle or ambiguous questions in a single retrieval round. Naive iterative retrieval can hurt rather than help: the first round already captures the best-matching segments, and subsequent undirected rounds tend to surface marginal or misleading content, causing the LLM to revise an initially correct prediction (Wang et al., 2024c; Tang et al., 2025). We address this with Contrastive Targeted Reflection (CTR), which activates when confidence falls below a threshold. Rather than retrieving more evidence indiscriminately, CTR prompts the LLM to articulate a discriminative gap, a statement of what is still missing to separate the top two candidate answers. An answer-anchoring constraint then requires any newly retrieved evidence to both support the revised prediction and contradict the original.

We integrate these three components into MSTree, a unified training-free framework. The main contributions are:

1. We present MSTree, a training-free framework for long-form VideoQA that builds on hierarchical retrieval to organize and retrieve sparse evidence through a multi-scale evidence tree.
2. We introduce a coordinated dual-modal segmentation strategy, in which visual anchor selection and text-guided dynamic programming

jointly produce visually and semantically coherent video segments.

3. We design Contrastive Targeted Reflection, an uncertainty-triggered reasoning mechanism that directs evidence retrieval and suppresses noise.
4. MSTree achieves state-of-the-art accuracy among training-free methods on EgoSchema, NExT-QA, and Video-MME, and improves further with stronger LLM backbones.

2 Related Work

Visual feature representation. Visual representations underpin modern video understanding. Built on the Vision Transformer (Dosovitskiy et al., 2020), CLIP (Radford et al., 2021) pioneered large-scale contrastive image-text pre-training and has become the de-facto starting point, with EVA-CLIP (Sun et al., 2023) scaling its training to deliver stronger visual features. BLIP (Li et al., 2022) and Flamingo (Alayrac et al., 2022) further bridge vision encoders with language models for cross-modal pre-training, and LLaVA (Li et al., 2024) adds visual instruction tuning. Recent open-source multimodal large language models (MLLMs), such as Qwen3-VL (Bai et al., 2025), InternVL 3.5 (Wang et al., 2025a), and Tarsier (Wang et al., 2024a), together with proprietary backbones (Comanici et al., 2025), handle high-resolution images and minute-scale videos

within a single model. Their effective context on hour-long videos, however, remains limited. Spatiotemporal encoders (Park et al., 2024; Wang et al., 2024d) extend image-level features with efficient long-range temporal modeling. HierVL (Ashutosh et al., 2023) and VideoReCap (Islam et al., 2024) construct multi-granularity video-language hierarchies, but these are produced offline at training time. MSTree adopts generic CLIP-based visual features and LLM-generated frame-level captions as inputs to its coordinated dual-modal segmentation, yielding content-aware temporal boundaries at inference time without retraining.

Video question answering. Early VideoQA research developed specialized neural architectures for spatiotemporal dynamics and compositional reasoning (Ko et al., 2023; Li et al., 2023b; Liu et al., 2025a; Liang et al., 2024). As large language models matured, the dominant paradigm shifted toward training-free pipelines. Image-grid methods (Kim et al., 2024; Lin et al., 2023) feed densely sampled thumbnails to a VLM, but the visual context window grows quickly with video length and may miss short events. Caption-then-reason frameworks instead convert the video into text and delegate reasoning to an LLM: LLoVi (Zhang et al., 2024) pioneered this paradigm with a minimal two-stage baseline; LangRepo (Kahatapitiya et al., 2025), MoReVQA (Min et al., 2024), and BOLT (Liu et al., 2025b) further refine it with multi-scale repositories, modular reasoning, and caption-driven prompts. These methods arrange captions as a flat stream, leaving the LLM to scan thousands of tokens with decisive cues diluted by surrounding context. Recent work introduces structure at different stages of video understanding. At the representation level, tree-based token merging compresses spatiotemporal redundancy and expands the usable video context without retraining (Fan et al., 2026). At the reasoning level, language-centric trees decompose complex questions into simpler perceptual subproblems and aggregate their answers along explicit reasoning paths (Liao et al., 2025). At the evidence level, planning, localization, retrieval, and adaptive sampling select relevant video content before final reasoning (Wang et al., 2024c; Yu et al., 2023; Ren et al., 2026; Tang et al., 2025). Caption-based hierarchies further support coarse-to-fine evidence access (Wang et al., 2025b); building on this direction, MSTree focuses on hierarchy construction and uses coordinated dual-modal segmentation to account for temporal continuity and cross-modal

semantic coherence.

Large language model based enhancement reasoning. Large language models support enhanced reasoning through structured prompting. Techniques such as Chain-of-Thought (CoT) prompting and information retelling (Li et al., 2025) have proven effective in refining imperfect video understanding by guiding the LLM to structure intermediate steps and supplement missing details. Verbal reinforcement (Shinn et al., 2023) is a core mechanism for iterative LLM self-correction. Extending these ideas to video, recent frameworks (Wang et al., 2023; Dang et al., 2025) apply iterative retrieval for complex temporal reasoning. These agentic approaches often rely on undirected retrieval, which risks surfacing distracting evidence and overriding correct initial predictions. MSTree addresses this through Contrastive Targeted Reflection (CTR), which conditions every retrieval step on an explicit discriminative gap.

3 Method

As illustrated in Figure 2, MSTree has three components: multi-scale evidence tree construction from coordinated visual and textual cues, coarse-to-fine LLM reasoning over the tree, and Contrastive Targeted Reflection.

3.1 Multi-Scale Evidence Tree Construction

Coordinated Dual-Modal Segmentation. Given a video, we uniformly sample n frames $\mathcal{F} = \{F_1, F_2, \dots, F_n\}$ and apply a pretrained vision-language model to generate a temporally ordered caption sequence $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$. The key challenge in tree construction is to partition $\mathcal{F}$ into segments that are internally coherent and temporally contiguous. Visual clustering ignores temporal order, while text truncation is blind to scene structure. The two modalities also carry complementary signals: visual anchors provide representative scene-level references, while captions characterize the narrative content around them, so optimizing only one of them discards the cross-modal alignment that defines a coherent event. We therefore use a two-stage segmentation method that coordinates the two modalities under contiguity constraints: visual cues set the coarse scene skeleton (Eq. 1), and textual cues refine the exact boundary of each scene within it (Eq. 2).

Visual Anchor Selection. For each frame, we extract a visual feature $\mathbf{v}_i = \phi_v(F_i)$ using a pre-

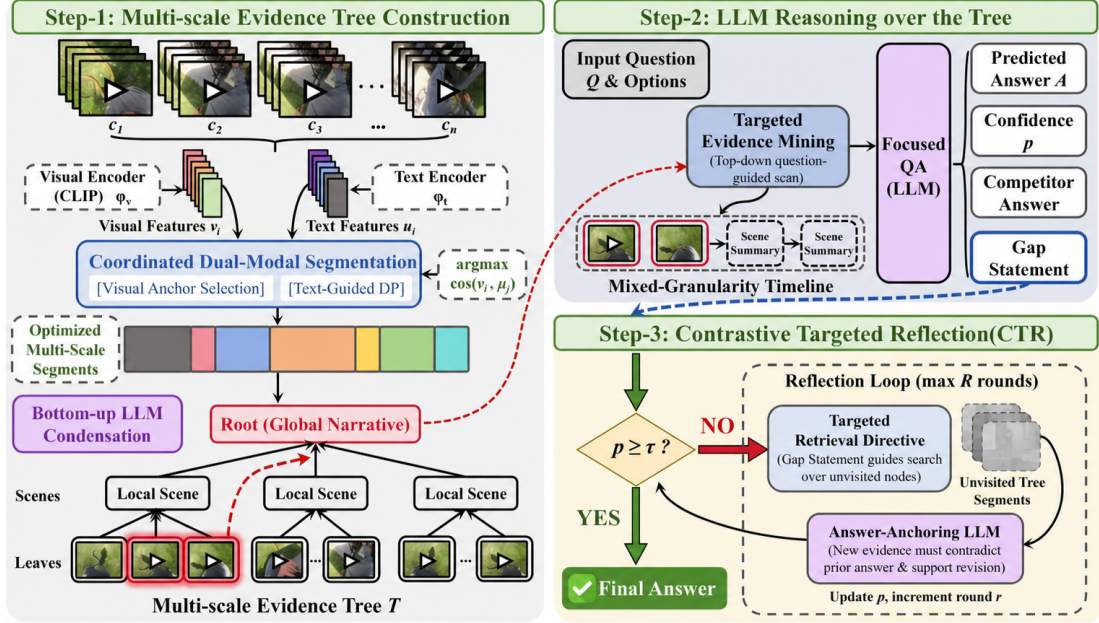


Figure 2: Overview of MSTree. Given a long video, coordinated dual-modal segmentation organizes frames into a Multi-Scale Evidence Tree. An LLM first mines the most relevant coarse segment, then reasons over its fine-grained captions. When confidence is insufficient, Contrastive Targeted Reflection retrieves targeted supplementary evidence.

trained visual encoder ϕ_v (e.g., EVA-CLIP (Sun et al., 2023)). Given a target segment count k , we initialize k intervals by uniform division and select the most representative anchor frame a_j within each interval:

$$a_j = \arg \max_{i \in \text{seg}_j} \cos(\mathbf{v}_i, \boldsymbol{\mu}_j), \quad (1)$$

where $\boldsymbol{\mu}_j$ is the normalized mean of all frame features in interval j . Each anchor locates the visual center of its semantic block and serves as a reference point for the subsequent boundary refinement.

Text-Guided Dynamic Programming Segmentation. We encode each caption with a text encoder ϕ_t (e.g., the CLIP ViT-L/14 text encoder (Radford et al., 2021)) to obtain $\mathbf{u}_i = \phi_t(c_i)$. The semantic prototype of segment j is defined as the mean text feature of the w frames surrounding anchor a_j , denoted $\mathbf{p}_j$. Segment boundaries are then optimized by minimizing:

$$\min_{\{[l_j, r_j]\}_{j=1}^k} \sum_{j=1}^k \sum_{i=l_j}^{r_j} (1 - \cos(\mathbf{u}_i, \mathbf{p}_j)), \quad (2)$$

subject to three constraints: (i) segments must be temporally contiguous; (ii) each anchor a_j must fall within its assigned segment range $[l_j, r_j]$;

and (iii) each segment must contain at least $L_{\min}$ frames. This optimization is solved exactly via dynamic programming (DP), and the resulting partition forms the bottom level of the multi-scale tree built next.

Hierarchical Summarization. We apply the segmentation algorithm recursively across multiple temporal scales to produce a nested, multi-granularity partition, from which we construct the multi-scale evidence tree $\mathcal{T}$. Leaf nodes correspond to individual frame captions. Each internal node stores an event summary derived from its children via bottom-up aggregation: an LLM first condenses the frame-level captions of a leaf cluster into a concise event description, which is then aggregated level by level into progressively broader scene overviews until the root node covers the full video. The resulting tree $\mathcal{T}$ encodes a unified semantic representation that serves as a structured index for on-demand evidence retrieval.

3.2 LLM Reasoning over the Evidence Tree

Given the multi-scale evidence tree $\mathcal{T}$ and a multiple-choice question Q , the reasoning process proceeds in two stages: targeted evidence mining followed by focused question answering.

Targeted Evidence Mining. Using $\mathcal{T}$ as a hierar-

chical semantic index, the LLM examines segment summaries against the question and candidate options, and selects the fine-grained segments most likely to carry the required evidence.

Focused Question Answering. The selected nodes are then expanded to their finest granularity, namely frame-level captions in $\mathcal{T}$, while unelected segments remain as coarse summaries to provide global temporal context. Over this mixed-granularity timeline, the LLM outputs a predicted answer, a confidence score p , the best competing candidate, and a discriminative gap statement specifying the evidence needed to distinguish the two candidates in the subsequent reflection stage.

3.3 Contrastive Targeted Reflection (CTR)

When the confidence score p meets or exceeds the threshold τ , the prediction from the previous stage is returned directly. For questions involving subtle action comparisons or partially overlapping option semantics, the evidence gathered in a single round may be insufficient: the first round commits to the segment that best matches the question’s surface description, but the segment most useful for ruling out a competing option often lies elsewhere in the video. Undirected iterative retrieval often surfaces marginally relevant segments that introduce noise, causing the LLM to overwrite a correct initial prediction (Wang et al., 2024c; Dang et al., 2025). CTR is triggered when $p < \tau$ to address this, replacing undirected retrieval with a contrastive, gap-driven query.

CTR uses the discriminative gap statement as a directive that guides a new round of evidence mining over unvisited segments. In the subsequent reasoning step, the LLM receives the accumulated evidence together with the newly retrieved segments. CTR also imposes an answer-anchoring constraint: any answer revision must be supported by new evidence that affirms the revised answer and contradicts the original one. This constraint reduces the risk that incidental noise overturns an otherwise correct initial prediction. CTR iterates for at most R rounds or until the confidence threshold is reached.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate MSTree on three long-form VideoQA benchmarks with complementary characteristics.

EgoSchema (Mangalam et al., 2023) comprises

5,000 multiple-choice QA pairs over 250+ hours of first-person footage, with each question grounded in a continuous three-minute clip. We report results on both the 500-question subset (Ego Sub.) and the full 5,000-question collection (Ego Full).

NExT-QA (Xiao et al., 2021) contains roughly 52K questions over 5,440 videos (44s avg. duration), partitioned into Temporal (Tem.), Causal (Cau.), and Descriptive (Des.) categories. We evaluate on the validation set and report per-category and overall (Avg.) accuracy.

Video-MME (Fu et al., 2025) is a long-form video understanding benchmark with 900 videos and 2,700 manually annotated multiple-choice questions across six visual domains. Video durations range up to one hour, with the long-video split averaging 44 minutes per video, substantially longer than the three-minute clips in EgoSchema and the 44-second segments in NExT-QA. This setting evaluates MSTree on hour-long content. We evaluate under the without-subtitles setting and report overall accuracy.

4.2 Implementation Details

Visual features are extracted with EVA-CLIP (Sun et al., 2023), and caption features are encoded with the CLIP ViT-L/14 text encoder (Radford et al., 2021). Frame-level captions are generated by LaViLa (Zhao et al., 2023) for EgoSchema and Video-MME, and by the LLoVi (Zhang et al., 2024) captioning pipeline for NExT-QA. Tree construction employs Mistral-7B (Jiang et al., 2023) for offline summarization, while inference uses GPT-4 (Achiam et al., 2023) as the primary LLM, with GPT-5 (Singh et al., 2025) additionally evaluated for backbone scalability. The schema $[1, 6, 36]$ sets the number of segments at each scale; the text prototype window w is set to 2 around each anchor and the minimum segment length to $L_{\min} = 2$. CTR uses a confidence threshold $\tau = 85$ with maximum reflection rounds $R = 2$.

4.3 Comparison with State-of-the-Art

We compare representative training-free baselines on EgoSchema and NExT-QA in Table 1 and on Video-MME in Table 2, grouped by inference-backbone family.

Main Results. Under the matched open-source-captioner/proprietary-LLM setting, MSTree achieves the highest accuracy on EgoSchema and NExT-QA. On EgoSchema, MSTree (GPT-4) reaches 68.4%/63.0% on Sub./Full, exceeding

Table 1: Comparison with state-of-the-art training-free methods on EgoSchema and NExT-QA. Best results in each group are underlined; overall best are in **bold**.

Method	(M)LLM	EgoSchema		NExT-QA			
		Sub.	Full	Tem.	Cau.	Des.	Avg.
<i>Open-source Captioners + Open-source LLMs</i>							
MVU (Ranasinghe et al., 2025)	Mistral-13B	60.3	37.6	55.4	48.1	64.1	55.2
LangRepo (Kahatapitiya et al., 2025)	Mixtral-8×7B	66.2	41.2	51.4	64.4	69.1	60.9
Video-LLaVA+INTP (Lin et al., 2024; Shang et al., 2024)	Vicuna-7B	—	38.6	58.6	61.9	72.2	62.7
<i>Proprietary MLLMs</i>							
IG-VLM (Kim et al., 2024)	GPT-4V	59.8	—	63.6	69.8	74.7	68.6
LVNet (Park et al., 2026)	GPT-4o	<u>68.2</u>	<u>61.1</u>	<u>65.5</u>	<u>75.0</u>	<u>81.5</u>	<u>72.9</u>
<i>Open-source Captioners + Proprietary LLMs</i>							
ProViQ (Choudhury et al., 2024)	GPT-3.5	57.1	—	—	—	—	64.6
MoReVQA (Min et al., 2024)	PaLM-2	—	51.7	64.6	70.2	—	69.2
Vamos (Wang et al., 2024b)	GPT-4	51.2	48.3	—	—	—	—
LLoVi (Zhang et al., 2024)	GPT-4	61.2	—	61.0	69.5	75.6	67.7
VideoAgent (Wang et al., 2024c)	GPT-4	60.2	54.1	64.5	72.7	81.1	71.3
VideoAgent (Fan et al., 2024)	GPT-4	62.8	60.2	—	—	—	—
LifelongMemory (Wang et al., 2023)	GPT-4	64.1	58.6	—	—	—	—
VideoTree (Wang et al., 2025b)	GPT-4	<u>66.2</u>	<u>61.1</u>	<u>70.6</u>	<u>76.5</u>	<u>83.9</u>	<u>75.6</u>
MSTree (Ours)	GPT-4	68.4	63.0	71.3	78.5	85.0	77.5
MSTree (Ours)	GPT-5	69.0	63.8	71.8	79.0	86.2	78.2

Table 2: Results on Video-MME (without subtitles). Best result in each group is underlined; overall best in **bold**.

Method	(M)LLM	Acc. (%)
<i>Open-source MLLMs</i>		
Video-LLaVA (Lin et al., 2024)	Vicuna-7B	39.9
VideoLLaMA-2 (Cheng et al., 2024)	Llama2-7B	<u>47.9</u>
<i>Open-source Captioners + Proprietary LLMs</i>		
LLoVi (Zhang et al., 2024)	GPT-4	52.3
VideoAgent (Wang et al., 2024c)	GPT-4	53.5
VideoTree (Wang et al., 2025b)	GPT-4	<u>54.7</u>
MSTree (Ours)	GPT-4	55.7
MSTree (Ours)	GPT-5	56.5

VideoTree by 2.2/1.9 points. It is comparable to GPT-4o-based LVNet on Sub. and exceeds it by 1.9 points on Full. On NExT-QA, MSTree improves over VideoTree by 1.9 points on average and across all question types. Section 4.4 analyzes these gains.

Long-Form Generalization on Video-MME.

Whereas EgoSchema and NExT-QA bound each question within a short clip, Video-MME tests understanding over longer videos that extend up to one hour. As shown in Table 2, MSTree (GPT-4) reaches 55.7% overall accuracy and ranks first among caption-based training-free methods, surpassing VideoTree by 1.0 point and LLoVi by 3.4 points. MSTree also outperforms several open-source MLLMs trained on large-scale video data, for example by 7.8 points over VideoLLaMA-2.

Although large proprietary long-context MLLMs still lead, the result shows that hierarchical evidence organization is effective on hour-long videos without any video-specific training. The gain over VideoTree persists on Video-MME, showing that MSTree remains competitive as video duration increases.

Backbone Scalability. Using GPT-5 (Singh et al., 2025) yields gains of 0.6, 0.7, and 0.8 points on EgoSchema Sub., NExT-QA Avg., and Video-MME, respectively, showing that MSTree continues to benefit from stronger reasoning backbones.

4.4 Ablation Studies

Component Contribution. Table 3 shows the effect of removing each major component. The largest drop (5.2 points) comes from replacing temporal segmentation with random equal-length splitting, indicating that semantically coherent segment boundaries, rather than the tree structure alone, are the primary driver of effective evidence retrieval. Removing the tree structure and reverting to flat retrieval reduces accuracy by 3.6 points. Removing CTR causes a more modest 0.8-point drop, consistent with its role as a complement to the structural improvements.

Segmentation Strategy and Tree Architecture.

Table 5 reports two studies on a fixed inference pipeline. For segmentation, random equal-length splitting gives a floor of 63.2%. Among single-

Table 3: Component ablation on EgoSchema Subset. Each row removes one component from the full system.

Configuration	Acc. (%)	Δ
MSTree (full system)	68.4	—
w/o temporal segmentation (random split)	63.2	−5.2
w/o tree structure (flat retrieval)	64.8	−3.6
w/o reflection mechanism ($R = 0$)	67.6	−0.8

modal variants, visual-only segmentation reaches 66.2% while text-only reaches 63.8%, suggesting that segment boundaries are mainly carried by visual feature changes. Combining both modalities outperforms either single-modal variant. Within the dual-modal variants, reversing the roles of the two modalities (text-guided anchor + visual DP) reaches 67.0%, 1.4 points below our configuration at 68.4%. This indicates that the two roles are not interchangeable: visual anchors provide representative scene-level references and are better suited to establishing the coarse scene skeleton, while textual features describe the narrative within a scene and are better suited to refining its boundaries. For tree architectures, all three tested schemas outperform the flat-retrieval variant in Table 3 (64.8%), showing that the benefit of hierarchical organization is retained across granularities. Their 2.8-point spread nevertheless indicates moderate sensitivity to segmentation granularity. Among them, the intermediate schema [1, 6, 36] strikes the best balance: the coarser [1, 4, 24] over-merges three-minute clips and hides local evidence, while the finer [1, 8, 48] leaves intermediate nodes too sparse to reliably guide the mining stage.

CTR Mechanism and Reflection Rounds. Table 6 compares three reflection strategies under the same backbone. Without reflection ($R=0$), the model reaches 67.6%. Naive iterative reflection drops accuracy to 67.0%, since undirected revision overwrites correct answers with tangential content. CTR raises accuracy to 68.4% (a 1.4-point gain over the naive approach) by anchoring each revision on a discriminative gap statement and requiring new evidence to actively contradict the original prediction.

Table 7 further sweeps the confidence threshold τ and maximum reflection rounds R in a 6×4 grid. Along the reflection-round axis (fixing $\tau=85$), accuracy reaches 68.4% for both $R=1$ and $R=2$, but drops to 67.2% at $R=3$, indicating that a small reflection cap is sufficient while further rounds tend to surface tangential content. Along the thresh-

Table 4: Per-question answer transitions along the reflection-round axis ($\tau = 85$). Each row is measured against the preceding setting: $W \rightarrow C$ denotes questions turned from incorrect to correct and $C \rightarrow W$ the reverse, both as percentages of all questions; Net is in points.

Max R	Acc. (%)	$W \rightarrow C$	$C \rightarrow W$	Net
0	67.6	—	—	—
1	68.4	1.6	0.8	+0.8
2	68.4	0.6	0.6	0.0
3	67.2	0.4	1.6	−1.2

Table 5: Ablation of segmentation strategy (top) and tree architecture (bottom) on EgoSchema Subset.

Variant	Acc. (%)
<i>Segmentation strategy</i>	
Random	63.2
Text-only	63.8
Visual-only	66.2
Text-guided anchor + visual DP	67.0
Visual anchor + text-guided DP (Ours)	68.4
<i>Tree architecture</i>	
[1, 4, 24]	65.6
[1, 6, 36] (Ours)	68.4
[1, 8, 48]	67.6

old axis (fixing $R=2$), $\tau=85$ attains the optimum 68.4%, while overly strict ($\tau=95$) or overly lenient ($\tau=65$) settings yield 68.2% and 67.2%, respectively. Across the full grid, 17 of the 24 configurations are within 1.0 point of the maximum; configurations within $\tau \in [75, 85]$ and $R \in \{1, 2\}$ are all within 0.8 points. These results indicate that CTR remains stable around the default configuration, with more consistent degradation under larger reflection limits.

Table 4 decomposes each accuracy change along the reflection-round axis into the questions that transition between correct and incorrect. The first round produces a net gain of 0.8 points through 1.6% corrections against 0.8% regressions. The second round produces equal 0.6% corrections and regressions, maintaining 68.4%; the third reverses the balance (0.4% corrections vs. 1.6% regressions) and lowers accuracy by 1.2 points. Thus, most gains arise in the first round, while extending beyond two rounds causes more harm than correction.

Robustness to Caption Quality. Like other caption-based methods, MSTree depends on caption quality. We examine the system’s robustness under caption corruption. On EgoSchema Subset we apply two perturbations: *random dropout* replaces 45% of all captions with an uninformative

Table 6: Reflection strategy ablation on EgoSchema Subset. The CTR row uses the default $\tau = 85$, $R = 2$.

Reflection Strategy	Acc. (%)
No reflection ($R = 0$)	67.6
Naive iterative reflection	67.0
CTR (Ours)	68.4

Table 7: CTR sensitivity to confidence threshold τ and maximum reflection rounds R on EgoSchema Subset. Default configuration ($\tau = 85$, $R = 2$) marked in **bold**.

τ	Reflection rounds R			
	$R=0$	$R=1$	$R=2$	$R=3$
95	67.8	67.4	68.2	67.0
90	67.8	68.4	66.8	66.8
85	67.6	68.4	68.4	67.2
80	68.0	68.0	67.6	66.6
75	67.4	67.8	68.2	67.4
65	67.6	67.8	67.2	66.6

placeholder, and *evidence noise* selectively replaces 45% of answer-supporting captions with nearby non-evidence captions matched in length and style.

As shown in Table 8, evidence noise causes a 2.4-point drop, whereas random dropout causes a smaller 0.6-point drop. The latter indicates limited sensitivity to random caption loss in this setting, while the larger drop under evidence noise shows greater sensitivity to corruption of answer-supporting captions. Caption quality therefore remains important, particularly when errors affect decisive evidence. Since the captioner is swappable, stronger captioning models may reduce this dependency as they become available.

4.5 Evidence Localization

Evidence Coverage. On EgoSchema Subset we annotate, for each question, the frame captions carrying the evidence needed to select the correct option, and group captions describing the same event into evidence groups. *Caption Recall* is the proportion of annotated evidence captions retrieved; *Group Recall* is the proportion of annotated evidence groups with at least one supporting caption retrieved. These annotations are used solely for di-

Table 8: Caption perturbation on EgoSchema Subset.

Setting	Acc. (%)	Δ
Original captions	68.4	—
45% captions $\rightarrow$ placeholder	67.8	-0.6
45% evidence captions $\rightarrow$ noise	66.0	-2.4

Table 9: Evidence localization on EgoSchema Subset. QA accuracy is repeated from Tables 1 and 3 for reference.

Method	CR (%)	GR (%)	QA Acc.
VideoTree (Wang et al., 2025b)	24.27	80.33	66.2
MSTree no reflection ($R=0$)	27.69	88.25	67.6
MSTree (Ours)	31.42	90.93	68.4

agnostic evaluation and are unavailable to MSTree during inference.

With reflection disabled, we compare MSTree and VideoTree, focusing the analysis on their different tree constructions. MSTree improves Group Recall by 7.92 points, the largest difference in Table 9, reducing unretrieved evidence groups from 19.67% to 11.75% (a 40% relative reduction). This pattern is consistent with scene-aligned boundaries keeping an event within a single segment, allowing evidence mining to access it as a unit rather than as fragments spread across segments. CTR then contributes 3.73 points of Caption Recall and 2.68 of Group Recall, consistent with gap-directed retrieval that both densifies evidence in a located event and reroutes mining to events missed in the first round. MSTree covers 90.93% of the annotated evidence groups while retrieving fewer than one-third of the annotated evidence captions. This pattern suggests that the evidence tree favors broad event-level coverage over exhaustive caption retrieval, consistent with selective coarse-to-fine evidence access. Across VideoTree, MSTree without CTR, and full MSTree, localization coverage rises with accuracy (66.2%, 67.6%, and 68.4%), showing that higher coverage accompanies the observed gains.

4.6 Qualitative Analysis

Figure 3 presents a case study from EgoSchema. In Round 1, the LLM matches the question against the L2 segment summaries and selects Seg5 (board-painting) as the segment most likely to carry option-discriminative evidence; focused QA over Seg5’s fine-grained captions then predicts Option A at 72% confidence. Since $72\% < \tau=85$, CTR is triggered and articulates the missing piece as a discriminative gap: “Option C requires laptop interaction evidence, not observed yet.” Round 2 mining acts on this gap and reroutes to Seg4, the only segment containing laptop frames, where focused QA over frames 75–76 supplies the decisive evidence and corrects the prediction to Option C at 90%. CTR thus converts Round-1 uncertainty into a targeted

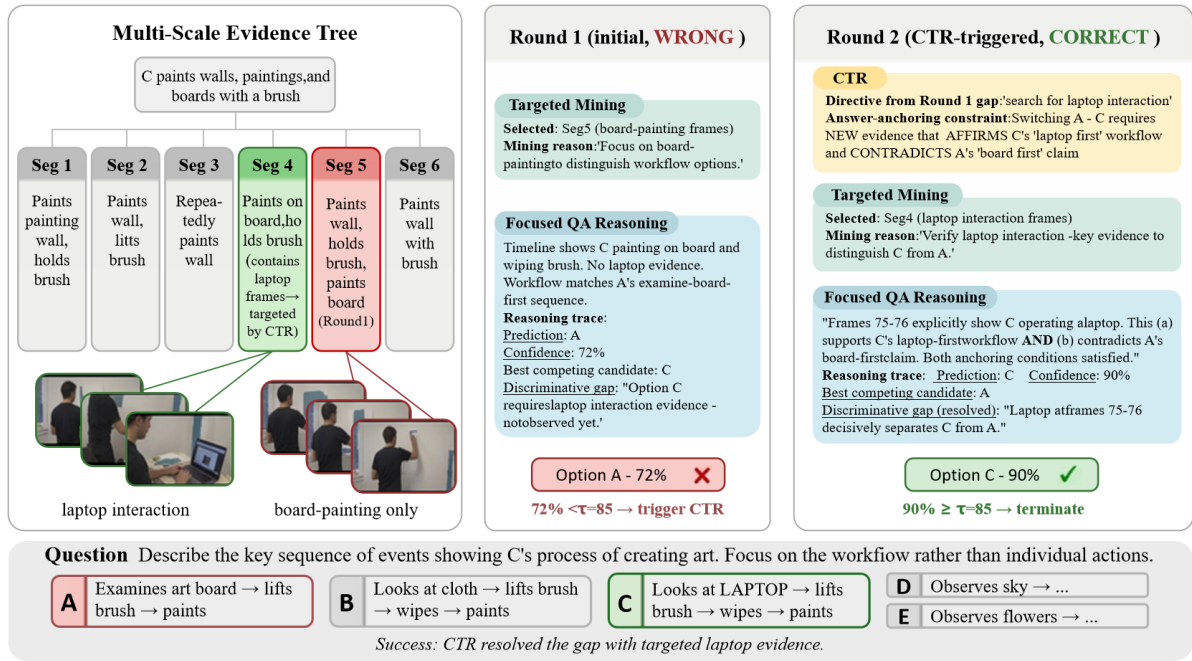


Figure 3: Qualitative case study on EgoSchema. A low-confidence Round 1 prediction (Option A, 72%) triggers CTR, whose discriminative gap reroutes mining from Seg5 to laptop evidence in Seg4 and corrects the answer to Option C at 90%.

directive, shifting mining from uninformative segments to the missing evidence in Seg4. The gap statement records what is missing from Round 1 and can be checked against the newly retrieved evidence in Round 2, making the resulting answer revision traceable to a specific evidence requirement.

Failure Modes. CTR remains dependent on the quality of the discriminative gap and the retrieved evidence. If either misses the decisive event, additional reflection rounds do not guarantee recovery and may introduce tangential evidence; surface-level token overlap may also satisfy the answer-anchoring constraint without event-level semantic support, leading to over-correction. Appendix E and Figure 4 provide a concrete failure case.

5 Conclusion

We present MSTree, a training-free framework for long-form VideoQA. Its Multi-Scale Evidence Tree is constructed through coordinated dual-modal segmentation, which integrates visual anchoring with text-guided boundary refinement to support a coarse-to-fine mine-then-reason workflow. Triggered only for low-confidence predictions, CTR uses a discriminative gap and an answer-anchoring constraint to guide targeted retrieval and reduce the risk of noise-induced answer revisions. Ex-

periments on egocentric and general long-form benchmarks show consistent gains over strong training-free baselines and further improvements with stronger backbones. Evidence-localization analysis further quantifies evidence coverage, while sensitivity results characterize variation across tree schemas and reflection settings. The framework cleanly separates content-driven indexing from query-driven retrieval, allowing its captioner, summarizer, and reasoning LLM to be upgraded independently.

6 Limitations

Like all caption-based VideoQA systems, MSTree is bounded by the captioner’s accuracy on the underlying frames. Systematic captioning errors propagate through the evidence tree into mining and focused QA, and coordinated dual-modal segmentation cannot recover from uniformly imprecise descriptions within a segment. This dependence is stronger when corruption affects decisive evidence than under random caption loss, and MSTree cannot independently verify caption accuracy. MSTree is modular by design: the captioner is swappable, so a stronger captioner can be substituted without retraining as such models improve. A second limitation arises from CTR’s text-level contrastive formulation. The discriminative gap and answer-

anchoring constraint specify what evidence should distinguish competing answers, but do not explicitly verify that the retrieved segment instantiates the corresponding event semantics. Consequently, lexical overlap across different events may satisfy the constraint and trigger an unsupported revision, while additional reflection rounds may reinforce an initially misdirected search. A promising direction is to couple CTR with stronger spatiotemporal grounding that aligns the entities, actions, and temporal relations expressed in the gap with specific video events before accepting an answer revision.

Acknowledgments

This work was supported by the 2025 Humanities and Social Sciences Research Planning Fund of the Ministry of Education of China (Grant No. 25YJA880062).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.
- Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. 2023. Hiervl: Learning hierarchical video-language embeddings. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23066–23078. IEEE.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.
- Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.
- Rohan Choudhury, Koichiro Niinuma, Kris M Kitani, and László A Jeni. 2024. Video question answering with procedural programs. In *European Conference on Computer Vision*, pages 315–332. Springer.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Jisheng Dang, Huilin Song, Junbin Xiao, Bimei Wang, Han Peng, Haoxuan Li, Xun Yang, Meng Wang, and Tat-Seng Chua. 2025. Mupa: Towards multi-path agentic reasoning for grounded video question answering. *arXiv preprint arXiv:2506.18071*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer.
- Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. 2026. Flashvid: Efficient video large language models via training-free tree-based spatiotemporal token merging. *arXiv preprint arXiv:2602.08024*.
- Chaoyou Fu, Yuhao Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118. IEEE.
- Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18198–18208. IEEE.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. *Mistral 7B*. *arXiv preprint arXiv:2310.06825*.
- Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. 2025. Language repository for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5627–5646.

- Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. 2024. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access*, 12:193057–193075.
- Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023a. IntentQA: Context-aware video intent reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11963–11974. IEEE.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Mingxin Li, Wenhao Wang, Hongru Ji, Xianghua Li, and Chao Gao. 2025. Can retelling have adequate information for reasoning? an enhancement method for imperfect video understanding with large language model. In *IJCAI*, pages 8150–8158.
- Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023b. Discovering spatio-temporal rationales for video question answering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13823–13832. IEEE.
- Tianming Liang, Chaolei Tan, Beihao Xia, Wei-Shi Zheng, and Jian-Fang Hu. 2024. Ranking distillation for open-ended video question answering with insufficient labels. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13161–13170. IEEE.
- Zhaohe Liao, Jiangtong Li, Siyu Sun, Qingyang Liu, Fengshun Xiao, Tianjiao Li, Qiang Zhang, Guang Chen, Li Niu, Changjun Jiang, and Liqing Zhang. 2025. Divide and conquer: Exploring language-centric tree reasoning for video question-answering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 37511–37527. PMLR.
- Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.
- Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. 2023. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*.
- Huabin Liu, Filip Ilievski, and Cees GM Snoek. 2025a. Commonsense video question answering through video-grounded entailment tree reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3262–3271. IEEE.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.
- Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. 2025b. Bolt: Boost large vision-language model without training for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3327. IEEE.
- Kartikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. EgoSchema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.
- Juhong Min, Shyamal Buch, Arsha Nagraani, Minsu Cho, and Cordelia Schmid. 2024. Morevqa: Exploring modular reasoning models for video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13235–13245. IEEE.
- Jinyoung Park, Hee-Seon Kim, Kangwook Ko, Minbeom Kim, and Changick Kim. 2024. Videomamba: Spatio-temporal selective state space model. In *European Conference on Computer Vision*, pages 1–18. Springer.
- Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S Ryoo. 2026. Too many frames, not all useful: Efficient strategies for long-form video qa. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3569–3588.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael Ryoo. 2025. Understanding long videos with multimodal language models. In *International Conference on Learning Representations*, volume 2025, pages 11810–11835.

- Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2026. Videorag: Retrieval-augmented generation with extreme long-context videos. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2390–2401.
- Yuzhang Shang, Bingxin Xu, Weitai Kang, Mu Cai, Yuheng Li, Zehao Wen, Zhen Dong, Kurt Keutzer, Yong Jae Lee, and Yan Yan. 2024. Interpolating video-llms: Toward longer-sequence llms in a training-free manner. *arXiv preprint arXiv:2409.12963*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.
- Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29118–29128. IEEE.
- Jiawei Wang, Liping Yuan, Yuchen Zhang, and Hao-miao Sun. 2024a. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*.
- Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. 2024b. Vamos: Versatile action models for video understanding. In *European Conference on Computer Vision*, pages 142–160. Springer.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025a. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.
- Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024c. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.
- Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. 2024d. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pages 396–416. Springer.
- Ying Wang, Yanlai Yang, and Mengye Ren. 2023. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*.
- Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025b. VideoTree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3272–3283. IEEE.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786. IEEE.
- Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771.
- Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737.
- Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6586–6597. IEEE.

A MStree Inference Procedure

The full inference procedure of MStree is summarized in Algorithm 1.

B Query-Agnostic Tree Construction

The coordinated dual-modal segmentation in MStree is query-agnostic: the tree is built from video content alone, so a single construction pass serves all questions about the same video. This amortizes the segmentation cost when one video is queried multiple times.

Question-driven boundary detection would instead require re-running segmentation per question, increasing latency and reducing throughput in repeated-query settings. The query-driven part of the framework is concentrated in the evidence mining and CTR stages, where the question determines which tree nodes are expanded to full granularity.

C IntentQA Experiments

To assess the cross-dataset generalization of MStree to a question distribution that emphasizes underlying intent rather than surface compositional reasoning, we additionally evaluate on IntentQA (Li et al., 2023a).

Dataset. IntentQA comprises approximately 16K multiple-choice QA pairs over 4,300 videos, derived from the NExT-QA video collection but deliberately reformulated to probe intent-centric reasoning rather than visual composition. We evaluate under the zero-shot setting and report overall accuracy on the test set.

Results. Table 10 summarizes results. MStree achieves 67.3% on IntentQA, surpassing the strongest training-free peer VideoTree (Wang et al., 2025b) (66.9%) and all other training-free baselines, while remaining competitive with the fine-tuned Vamos (Wang et al., 2024b) (68.5%). The margin over VideoTree is smaller than on EgoSchema and NExT-QA, which reflects the nature of intent-centric reasoning: intent often hinges on actor goals and contextual cues that span the entire video rather than concentrating in a single segment. MStree nonetheless leads VideoTree by 0.4 points, indicating that the structural advantages of coordinated dual-modal segmentation and CTR transfer beyond the question types originally targeted.

Table 10: Results on IntentQA test set. Training-free methods are compared against fine-tuned references for context.

Method	Acc. (%)
<i>Fine-tuned Methods</i>	
VGT	51.3
Vamos (Wang et al., 2024b)	68.5
<i>Training-free Methods</i>	
LangRepo (Kahatapitiya et al., 2025)	59.1
SeViLA (Yu et al., 2023)	60.9
LLoVi (Zhang et al., 2024)	64.0
IG-VLM (Kim et al., 2024)	64.2
VideoTree (Wang et al., 2025b)	66.9
MStree (Ours)	67.3

Table 11: Inference-budget comparison on the EgoSchema Subset, all using GPT-4 as the reasoning backbone. “Captions” is the average number of textual content units the LLM observes per question (segment summaries plus selected frame captions); “LLM Calls” includes the one-off Stage-0 summarization.

Method	Captions	LLM Calls	Acc. (%)
VideoAgent (Wang et al., 2024c)	11.0	9.0	57.4
VideoTree (Wang et al., 2025b)	11.3	2.8	62.2
MStree (Ours)	11.3	4.6	64.0

D Inference Budget Comparison

Table 11 reports MStree’s average caption count and LLM-call count alongside VideoAgent (Wang et al., 2024c) and VideoTree (Wang et al., 2025b) on the EgoSchema Subset under the same GPT-4 backbone. At a comparable caption budget (11.3 vs. 11.0 and 11.3), MStree achieves the highest accuracy at 64.0%, exceeding VideoAgent by 6.6 points and VideoTree by 1.8 points. Its 4.6 LLM calls per question are 49% fewer than VideoAgent’s 9.0 but 1.8 higher than VideoTree’s 2.8. This reflects a clear trade-off: MStree uses a moderately larger LLM-call budget than VideoTree and in return delivers higher accuracy. We consider this an acceptable allocation, since the additional LLM compute is invested in retrieving and reasoning over the supplementary evidence required to arrive at accurate answers.

E Failure Case Analysis

Figure 4 illustrates the principal failure mode of the reflection mechanism. In Round 1, the LLM mines Seg6 (“adjusts stone & soil, lifts brick mold”) and

Algorithm 1: MSTree inference procedure. Tree summarization is a one-off offline step; the reasoning LLM $\mathcal{M}$ is invoked at evidence mining and focused QA, with optional gap-guided reflection. Coordinated dual-modal segmentation at each scale selects visual anchors and then refines the boundaries by text-guided DP against the textual prototypes.

Require : Video $\mathcal{V}$, question Q , reasoning LLM $\mathcal{M}$, tree schema $[1, a, b]$, confidence threshold τ , max reflection rounds R

Result: Predicted answer $pred$

```

// Stage 1: Multi-Scale Evidence Tree construction
1 Sample  $n$  frames from  $\mathcal{V}$  and encode them, together with their captions, using a visual encoder  $\phi_v$ 
  and a text encoder  $\phi_t$ ;
2 Recursively partition the video at two scales (into  $a$  coarse and  $b$  fine-grained segments) by
  coordinated dual-modal segmentation: select visual anchors, then refine the boundaries by
  text-guided DP under temporal contiguity;
3 Aggregate captions bottom-up via an offline summarizer to obtain the multi-scale evidence tree  $\mathcal{T}$ ;

// Stage 2: Targeted Evidence Mining
4  $segs \leftarrow \mathcal{M}$  selects the fine-grained segments most relevant to  $Q$  from the segment summaries in  $\mathcal{T}$ ;
5  $ctx \leftarrow$  frame captions inside  $segs$ , plus coarse summaries of the unselected segments;

// Stage 3: Focused QA with Contrastive Targeted Reflection
6  $(pred, conf, gap) \leftarrow \mathcal{M}(ctx, Q)$ ;
7 for  $r \leftarrow 1$  to  $R$  do
8   if  $conf \geq \tau$  then break;
9    $pred_{prev} \leftarrow pred$ ;
10   $new \leftarrow$  additional fine-grained segments mined from unvisited summaries, guided by the gap
    statement  $gap$ ;
11   $ctx \leftarrow ctx \cup \{\text{frame captions inside } new\}$ ;
12   $(pred, conf, gap) \leftarrow \mathcal{M}(ctx, Q, pred_{prev})$ , revising only if new evidence supports  $pred$  and
    contradicts  $pred_{prev}$ ;
13 end
14 return  $pred$ ;

```

reasons over its fine-grained captions to correctly predict Option A (“build a brick wall”) at 75% confidence, with a well-formed gap stating that “Option D would require tree-removal evidence, not observed.” The outcome is borderline: the brick-mold and stone-moving evidence is concrete but not exclusively decisive, since incidental axe and spade actions in the same scene keep confidence below the $\tau=85$ threshold, so CTR activates on a correct but low-confidence answer.

In Round 2, the directive “find tree-removal evidence” is acted on by surface-level token matching: Seg2’s “hits spade with hammer” and Seg6’s tail action “throws axe” contain tool tokens (spade, axe) statistically associated with tree removal, even though both actions describe brick-mold preparation. The answer-anchoring constraint is satisfied only syntactically: the tokens superficially contradict “brick wall” and affirm “tree removal”, yet

the underlying scene semantics never support the switch, so the model over-corrects to Option D at 85%. This case localizes a concrete weakness: the answer-anchoring check operates on the gap’s surface predicate rather than grounded referents, so ambiguous tool nouns can be passed by spurious matches. A natural fix is to extend the anchor with explicit referent grounding, requiring the verifying segment to instantiate the gap’s noun in the same activity context, or a multi-step gap-verification stage that filters surface matches before they override the initial answer.

F Prompt Templates

This section presents the three core prompts used by MSTree at inference time, aligned with the three modules described in Sections 3.2–3.3: **Targeted Evidence Mining** (TEM, Table 12), **Focused Question Answering** (FQA, Table 13), and **Con-**

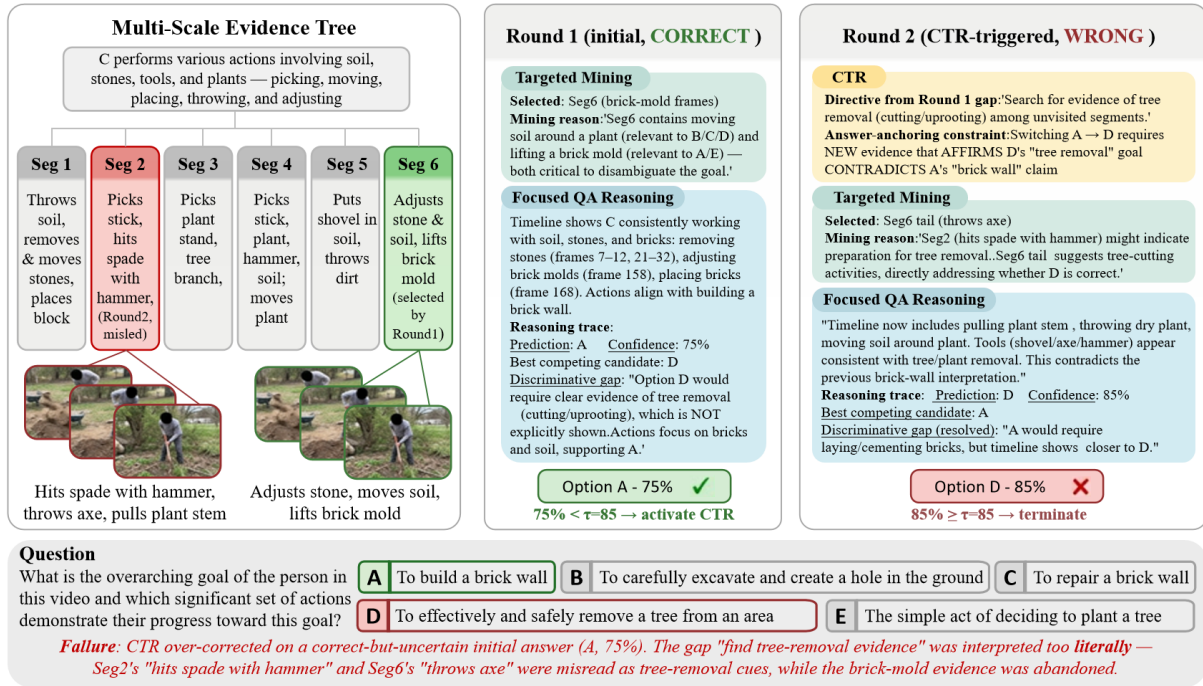


Figure 4: Failure case on EgoSchema. Round 1 mines Seg6 and correctly predicts Option A (“build a brick wall”) at 75% with the gap “Option D would require tree-removal evidence, not observed.” Since $75\% < \tau=85$, CTR still triggers. In Round 2 the directive is taken too literally: Seg2’s “hits spade with hammer” and Seg6’s tail action “throws axe” superficially match “tree removal” but describe brick-mold preparation, so the model over-corrects to Option D at 85%.

Table 12: Prompt template for Targeted Evidence Mining (TEM).

System Select fine-grained video segments whose detail is most critical for answering the question. Respond with JSON only.
User Which fine-grained segments need to be examined in detail to answer this question? Question: question (A) optionA (B) optionB (C) optionC (D) optionD (E) optionE [Video Overview (L1)] level1_summary [Coarse Segment Summaries (L2)] level2_summaries [Fine-Grained Segment Summaries (L3)] level3_summaries Select the fine-grained (L3) segments (by their ID) whose summaries are most critical for distinguishing the answer options. Respond JSON: {"selected_level3_ids": [...], "reason": "..."}
Assistant <pre> {"selected_level3_ids": [...], "reason": "one-sentence rationale citing L1/L2 evidence"} </pre>

trastive Targeted Reflection (CTR, Table 14).

Table 13: Prompt template for **Focused Question Answering (FQA)**.

System
Answer multiple-choice questions about videos using the provided timeline evidence. Respond with JSON only.
User
Question: <code>question</code> (A) <code>optionA</code> (B) <code>optionB</code> (C) <code>optionC</code> (D) <code>optionD</code> (E) <code>optionE</code> [Video Timeline Evidence] <code>flat_timeline</code> Evaluate each option against the timeline evidence. Choose the best-supported one. Identify the runner-up and what evidence would distinguish them. Respond JSON: {"prediction": "X", "confidence": 0-100, "reason": "...", "top2_rival": "Y", "discriminative_gap": "..."} Assistant {"prediction": "letter", "confidence": integer 0-100, "reason": "evidence-cited rationale", "top2_rival": "letter", "discriminative_gap": "fact that would distinguish prediction from rival"}

Table 14: Prompt template for **Contrastive Targeted Reflection (CTR)**.

User (appended to TEM prompt)
Same context as Table 12 (Question, Options, L1 / L2 / L3 blocks), plus: Already examined (do NOT reselect): <code>forbidden_ids</code> Previous unresolved uncertainty: " <code>prev_discriminative_gap</code> " Pick new fine-grained (L3) segments (by their ID) most likely to resolve this uncertainty. Respond JSON: {"selected_level3_ids": [...], "reason": "..."} Assistant {"selected_level3_ids": [...], "reason": "rationale citing prev gap and the newly resolving content"}