Caption-once, Frames-on-Demand: Visual-Need Routing for Budget-Aware Agentic Long Video Understanding

Weitong Cai[1], **Hang Zhang**[2*], **Yukai Huang**[3], **Yiqiao Xie**[4], **Shan Gao**[5], **Jiankang Deng**[4], **Songcen Xu**[5], **Jifei Song**[5], **Zhensong Zhang**[5*]
[1]Queen Mary University of London, [2]Independent Researcher,
[3]Durham University, [4]Imperial College London, [5]Huawei
weitong.cai@qmul.ac.uk, miruku.hzhang@gmail.com, zhangzhensong@huawei.com

Abstract

Long-video understanding on edge devices must reason over hours of content under tight compute and bandwidth budgets. Subsampling visual tokens loses temporal structure, while text-only video memories lose fine-grained visual attributes. We observe a visual-textual duality: language memories carry long-range temporal structure better than dense frames, while pixels remain decisive for attribute-level perception. Building on this insight, we propose **Caption-once, Frames-on-Demand (CFD)**, a budget-aware edge-cloud agentic framework. The edge runs a single offline captioning pass that builds a dual-track narrative index, an event-level story skeleton plus a clip-level micro-log, cached and reused across queries without re-captioning. At query time, a cloud-side MLLM reasons over the index in a story-first loop centered on a **lightweight Visual-Need Router**: a per-query gating module that triggers bounded keyframe retrieval only for perceptual questions (appearance, on-screen text, attribute disambiguation) and keeps temporal-structural questions in language space. The router turns visual access into a first-class, query-conditioned cost, capping per-query frame consumption regardless of video length. Experiments on long-video benchmarks demonstrate strong accuracy-efficiency trade-offs while substantially reducing online visual processing.

1 Introduction

Long-form video understanding underlies a growing class of multimodal applications, from interactive video assistants and on-device personal memory to wearable visual aids, each constrained by tight inference budgets imposed by either edge hardware or latency-sensitive cloud serving (Paruchuri et al., 2025; Singh and Gill, 2023). Under such constraints, long video QA becomes a joint modeling and resource-allocation problem: the system must preserve query-relevant memory over the entire video horizon, yet missing even a small but decisive event or attribute can directly cause incorrect answers (Mangalam et al., 2023).

Existing approaches can be viewed along two intertwined design axes: **(i) Visual compression** methods (fig. 1(a)) reduce cost by sampling fewer frames or compressing visual tokens, but may sacrifice coverage or fidelity on long videos (Qian et al., 2024; Jiang et al., 2025). **(ii) Textual translation** methods (fig. 1(b)) convert videos into captions, transcripts, or language memories and perform retrieval and reasoning in text space, or iteratively revisit video segments to produce updated textual memories across multiple rounds (Fan et al., 2024; Zhi et al., 2025; Zuo et al., 2025; Cai et al., 2025). These strategies improve scalability, but text-centric representations can lose fine-grained visual evidence (Wang et al., 2023; Kahatapitiya et al., 2025), while iterative revisiting often introduces query-time cost that grows with interaction rounds and is difficult to bound under strict edge budgets (Fan et al., 2024; Zhi et al., 2025).

In preliminary studies, we observe a consistent duality. For image-level referencing (Table 1), caption-only reasoning falls short of reasoning from original pixels, since fine-grained properties such as identity, color, and attribute binding are easily lost under aggressive linguistic compression (Paruchuri et al., 2025). In contrast, for long-horizon temporal reasoning (Table 2), language memories can sometimes be more effective than dense visual inputs for chronological understanding and event-level relations, because they provide a compact, discrete record of what happened and in what order (Wang et al., 2023; Kahatapitiya et al., 2025). This suggests that text is not merely a lossy substitute for video: it can be a stronger carrier for temporal structure, while sparse raw frames remain essential for verifiable visual evidence. The intu-

*Corresponding authors.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 9651–9676
October 24-29, 2026 ©2026 Association for Computational Linguistics

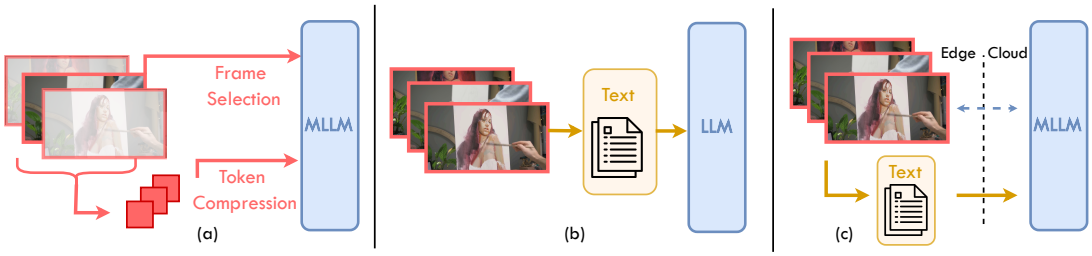

Figure 1: **Long video understanding. (a)** Visual compression methods (frame selection or token compression) reduce visual tokens fed to an MLLM but risk losing coverage on long videos. **(b)** Textual translation methods convert video into language memories for LLM reasoning, improving scalability but losing fine-grained visual evidence. **(c)** Our approach separates offline narrative indexing on the edge from on-demand visual verification on the cloud, combining textual temporal coverage with targeted pixel-level evidence only when needed.

ition aligns with human memory: we rarely retain every pixel-level detail, but we can retain and query a concise storyline of an experience.

Motivated by this observation, we propose a **Caption-once, Frames-on-Demand (CFD)** paradigm (fig. 1(c)) for budget-aware long video understanding in an edge-cloud setting. The key idea is to index temporal structure once in language and reuse it across queries, while retrieving raw frames only when fine-grained visual attribution is necessary. Under this view, efficiency does not come from uniformly reducing visual computation, but from allocating visual access selectively to attribute-critical moments. We instantiate this paradigm with CFD, a budget-aware edge-cloud agentic pipeline for long video understanding. On the edge, a lightweight MLLM performs *single-pass, question-agnostic indexing* of the full video stream to build two complementary timestamped language memories: an *event-level story skeleton* derived from shot/event segmentation (Soucek and Lokoc, 2024), and a *uniform clip-level micro-log* that preserves local temporal details between events. Both memory tracks are produced once per video and cached for reuse across subsequent queries without re-captioning the video.

At query time, a stronger cloud-side MLLM reasons over the cached narrative index in an iterative backtracking loop. It first attempts a *story-first* answer from the event memory. If confidence is insufficient, the agent *localizes* a relevant unexplored event, *enriches* it with overlapping clip-level captions, and re-attempts answering. A lightweight *Visual-Need Router* then predicts whether pixel-level inspection is required: it favors frame retrieval for perceptual queries (*e.g.*, appearance, on-screen text) and skips it for temporally grounded questions when narrative memory is sufficient. When visual inspection is triggered, the system retrieves a bounded set of keyframes from the localized event into a fixed-capacity FIFO working memory, ensuring that per-query visual cost remains controlled regardless of video length. The agent then performs *multimodal re-reasoning* over textual memory and retrieved visual evidence in a unified prompt. The loop repeats until confidence is reached or the query budget is exhausted.

In summary, our contributions are threefold:

- **Visual-Need Router.** We introduce a per-query gating module that classifies each question as perceptual or temporal-structural and triggers raw-frame retrieval only for the former. To our knowledge, this is the first agentic video-QA design that conditions visual access on question type, converting frame retrieval from an emergent pipeline byproduct into an *explicit, query-conditioned cost*.
- **Caption-once, Frames-on-Demand framework.** Building on prior caption-once paradigms (Zhang et al., 2024a; Kahatapitiya et al., 2025), we couple a reusable dual-track narrative index (event skeleton plus clip-level micro-log) with the Visual-Need Router and a fixed-capacity FIFO working memory, exposing per-query visual cost as an explicit hyperparameter rather than an implicit consequence of interaction depth.
- **Empirical study.** On Video-MME and InfiniBench, CFD reaches accuracy competitive with prior agent-based methods while using roughly an order of magnitude fewer frames per question, supporting the design hypothesis that visual access can be allocated selectively rather than uniformly.

An extended discussion of related work is provided in Appendix C.

Table 1: CaptionQA (Yang et al., 2025a). Caption:`Qwen2.5-VL-7B`; Reasoning:`Qwen3-VL-4B`.

Input	Overall (%)
Image	92.32
Caption	77.17

Table 2: InfiniBench (Ataallah et al., 2025). Use `Qwen3-VL-32B` for caption and reasoning. Results show captions have better performance on temporal context understanding than raw videos.

Pipeline	Chronological Understanding	Global Appearance	Scene Transitions	Character Actions
Video	48.44	70.54	53.97	67.04
60s / Caption	56.46	53.49	58.73	53.07

2 Method

We present Caption-once, Frames-on-Demand (CFD), a budget-aware edge-cloud agentic framework for long video understanding. The framework consists of three parts: a three-tier memory architecture that separates offline narrative indexing from online visual verification (section 2.1), four agent roles (section 2.2), and a story-first reasoning loop that orchestrates these agents under a controlled visual budget (section 2.3). An overview is shown in fig. 2.

2.1 Three-Tier Edge-Cloud Memory Architecture

A central challenge in long video understanding is that different questions demand different evidence. Temporal and causal questions (*e.g.*, "What happened after the man left?") can often be resolved from a compact narrative record, while perceptual questions (*e.g.*, "What color was the mug?") require direct pixel-level verification. Serving both needs efficiently calls for a memory that provides broad temporal coverage by default, fine-grained textual detail on demand, and targeted visual evidence only when language alone is insufficient.

CFD addresses this with a three-tier memory architecture under an edge-cloud setting. Prior agent-based systems (Zuo et al., 2025) adopt a hierarchical memory where deeper levels are generated *online* by repeatedly invoking the captioning model during question answering, incurring query-time visual cost that grows with interaction depth. In contrast, CFD constructs all textual memories *offline* in a single question-agnostic indexing pass, and introduces a third, *visual* tier that provides raw pixel evidence when text is insufficient. This decoupling ensures that query-time reasoning operates primarily in language space, with visual access as a bounded, selective supplement rather than a recurring expense. We emphasise that the framework does not presume the offline captions to be sufficient on their own: the visual tier is designed precisely for cases where the question-agnostic index lacks the needed detail, with the Visual-Need Router (section 2.2) deciding when to escalate from text to pixels rather than committing to a possibly wrong text-only answer.

Tier 1: Event memory (global story skeleton). Given a video V of duration D seconds, we segment it into semantically coherent events using a shot-boundary detector (Soucek and Lokoc, 2024). Adjacent segments shorter than a threshold $\tau_{\min}$ are merged into the preceding segment, yielding variable-length events $\{e_1, \ldots, e_K\}$. A lightweight captioning MLLM then processes each event to produce a structured narrative profile:

$$m_k^E = \text{Captioner}(e_k,\ p_{\text{event}}), \tag{1}$$

where a prompt p_{event} instructs the captioner to produce a story-skeleton description including scene overview, entity listing, chronological event flow, and retrieval anchor tags. The full event memory $\mathcal{M}_E = \{m_k^E\}_{k=1}^K$ spans the entire video and is *always presented in full* to the reasoning agent, serving as the global timeline and starting point for all question answering.

Tier 2: Clip memory (local temporal details). Independent of event segmentation, we partition V into fixed-length clips of duration τ_c with configurable stride, yielding a sequence of clips $\{c_1, \ldots, c_L\}$. Each clip is captioned with a micro-action-level prompt p_{clip} designed for high temporal granularity, one action per sentence:

$$m_j^C = \text{Captioner}(c_j,\ p_{\text{clip}}). \tag{2}$$

$\mathcal{M}_C = \{m_j^C\}_{j=1}^L$ captures local temporal details (micro-actions, state changes, visible text) that may be too fine-grained for the event-level skeleton.

Crucially, the clip memory is *not* fed to the reasoning agent in full. Instead, it is activated selectively: when the reasoning loop localizes a specific event e^*, all clip captions whose time intervals overlap with e^* are injected as nested sub-descriptions within e^*'s narrative text (section 2.3). This on-demand injection avoids overwhelming

the LLM context with full clip-level detail for the entire video, while providing dense local information exactly where it is needed. Moreover, these enrichments are *cumulative*: once clip captions are injected for an explored event, they persist in the narrative memory for all subsequent reasoning steps within the same query. This mechanism replaces the online re-captioning used in prior work (Zuo et al., 2025): instead of invoking the MLLM again to generate finer descriptions, CFD composes already-cached captions through pure text operations at zero additional visual cost.

Tier 3: Visual working memory (pixel-level evidence). The first two tiers operate entirely in language space. For questions that require direct visual verification, CFD maintains a Visual Working Memory $\mathcal{W}$, a FIFO queue with a deliberately small fixed capacity N_{wm} that stores raw image frames extracted from the video. When the reasoning loop determines that pixel-level inspection is necessary for an event e^*, keyframes are adaptively sampled from e^* and pushed into $\mathcal{W}$; when $\mathcal{W}$ is full, the oldest frames are evicted. The capacity N_{wm} is set to be orders of magnitude smaller than the total frame count of the video (e.g., 16 frames for a one-hour video). This compact design serves two purposes: it prevents excessive visual tokens from diluting the LLM's attention over the context, and it minimizes the volume of visual data transmitted from edge to cloud, keeping uplink bandwidth and energy consumption under control. By introducing a bounded visual channel alongside the narrative, CFD can ground perceptual claims in actual pixels while keeping the visual footprint minimal.

The three tiers form a coarse-to-fine progression that mirrors the caption-once, frames-on-demand principle: Event Memory provides breadth across the full video, Clip Memory adds depth at targeted time periods, and Visual Working Memory supplies pixel-level evidence when language alone is insufficient. All three tiers are jointly presented to the answering agent in a single multimodal prompt, with textual memories providing the reasoning context and visual evidence grounding fine-grained attributes.

2.2 Agent Roles in the CFD Framework

CFD employs four agent roles instantiated through prompt engineering, inspired by prior multi-agent designs for video understanding (Zuo et al., 2025; Zhi et al., 2025; Fan et al., 2024). A key difference from these systems is the *absence of an Instruction Agent*: prior pipelines use a dedicated agent to analyze what information is missing and generate re-captioning instructions, triggering additional VLM calls at query time (Zuo et al., 2025). CFD eliminates this role entirely by substituting it with two complementary mechanisms: the pre-computed clip memory (which already contains the finer-grained descriptions that an Instruction Agent would request) and a Visual-Need Router (which decides when to look at raw pixels instead of generating more text).

Captioning Agent (offline only). An edge-side MLLM processes each event clip and uniform clip according to eqs. (1) and (2). This agent is active *only during the offline indexing stage* on the edge. Once the dual-track memory is cached, the agent is not invoked again for any subsequent query on the same video.

Answering Agent. Given the current narrative memory $\mathcal{M}$ (which may include injected clip details) and the visual working memory $\mathcal{W}$, this agent evaluates whether the available evidence is sufficient to answer the question Q with confidence:

$$(a,\ \textit{conf},\ r) = \textsc{AnsAgent}(\mathcal{M},\ \mathcal{W},\ Q), \quad (3)$$

where a is the candidate answer, $\textit{conf} \in \{\textit{true}, \textit{false}\}$ is the confidence flag, and r is a chain-of-thought reason. The textual memory and visual evidence are assembled into a single multimodal prompt: images from $\mathcal{W}$ are interleaved with their timestamps before the textual narrative, allowing the LLM to jointly ground visual details and temporal context. When $\textit{conf} = \textit{true}$, the answer a is returned; otherwise, the system enters/continues the reasoning loop.

Localization Agent. This agent identifies the single most question-relevant event from the narrative memory, excluding previously explored events:

$$e^* = \textsc{LocAgent}(\mathcal{M}_E \setminus \mathcal{S}_{\text{tried}},\ \mathcal{W},\ Q), \quad (4)$$

where $\mathcal{S}_{\text{tried}}$ is the set of events already visited in the current query session. By operating over progressively narrowed candidate sets, the agent ensures that each reasoning iteration explores a new temporal region, expanding the breadth of evidence gathering.

Visual-Need Router. After an enriched text-only answer attempt fails to reach confidence, this agent decides whether pixel-level inspection of the selected event e^* is likely to improve correctness:

$$\textit{need_visual} = \textsc{Router}(e^*,\ Q,\ r), \quad (5)$$

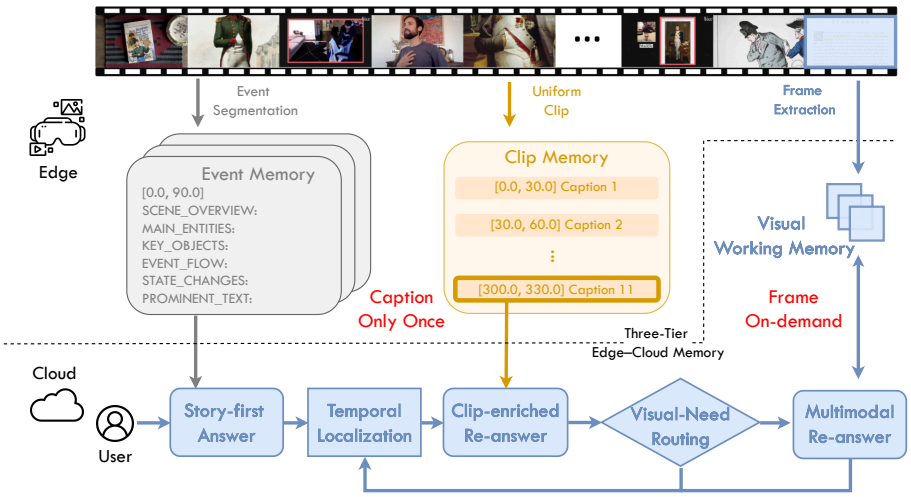

Figure 2: **Overview of CFD.** On the edge, a lightweight MLLM performs a single offline captioning pass to build a dual-track narrative index: an event-level story skeleton and a clip-level micro-log, both cached and reused across queries without re-captioning. At query time, a cloud-side MLLM reasons over the cached index through a story-first iterative loop, enriching targeted events with clip details and retrieving raw frames on demand only when the Visual-Need Router determines pixel-level evidence is necessary.

where r is the reason from the preceding answer attempt. The router follows a perceptual-versus-temporal rubric. It accepts (triggers frame extraction) when the question targets visual attributes such as appearance, on-screen text, spatial layout, or object disambiguation, i.e., evidence that cannot be reliably inferred from text alone. It declines (skips frame extraction) when the question primarily concerns temporal ordering, scene transitions, or long-range narrative structure, where the textual memory already provides a suitable abstraction. When declining, the router clears the visual working memory $\mathcal{W}$ and directs the loop to the next iteration without incurring any visual cost. This role is unique to CFD: existing agent-based video systems (Zuo et al., 2025; Wang et al., 2025d; Ma et al., 2025) do not gate visual access based on question type, and therefore either always or never use visual frames regardless of whether the question demands them.

Crucially, the router is *not* a downstream perception module but an explicit cost-control gate: it deliberately routes some questions away from frames in order to keep the per-query frame budget tight. This means the router may sacrifice marginal accuracy on attribute-centric questions in exchange for a substantially lower average visual footprint, a trade-off we view as a feature rather than a bug, because it exposes per-query visual cost as a tunable hyperparameter rather than an emergent byproduct of pipeline depth. Applications that prioritise attribute fidelity over visual cost can simply disable the router and always route to frames; we report both configurations in our ablations (section 3).

2.3 Story-First Reasoning Loop

At query time, the system receives a question Q and the pre-built memories $\mathcal{M}_E$ and $\mathcal{M}_C$. It initializes the visual working memory $\mathcal{W} = \varnothing$, the explored set $\mathcal{S}_{\text{tried}} = \varnothing$, and a cumulative clip enrichment bank $\mathcal{B} = \varnothing$ that tracks which clip captions have been injected into which events. The reasoning proceeds through the following steps, with a maximum budget of T iterations.

Step 1: Story-first answer. The Answering Agent receives the full Event Memory $\mathcal{M}_E$ and the empty $\mathcal{W}$, and attempts to answer Q. If confident, the answer is returned immediately. This early exit handles temporal and narrative questions that can be resolved from the story skeleton alone, requiring only a single LLM call for the entire query. Importantly, CFD does not only route between textual and visual evidence. Its story-first agentic loop also adaptively allocates the depth of reasoning and evidence access according to whether the currently available memory is sufficient.

Step 2: Temporal localization. The Localization Agent selects the most relevant event e^* from $\mathcal{M}_E \setminus \mathcal{S}_{\text{tried}}$, and e^* is added to $\mathcal{S}_{\text{tried}}$.

Step 3: Clip-enriched re-answer. All Clip Memory entries overlapping with e^* are collected and injected as nested sub-descriptions within e^*'s event

caption, and these enrichments are added to $\mathcal{B}$. The narrative memory is then rebuilt: for every event e_k, if $\mathcal{B}$ contains clip enrichments for e_k, the clip captions are appended chronologically under a "More details" block within e_k's text; the base event text is always restored first to prevent nesting accumulation. The Answering Agent re-attempts with this enriched memory. Because $\mathcal{B}$ accumulates across iterations, the narrative grows progressively richer as more events are explored, carrying forward all previously gathered local detail. If confident, the answer is returned.

Step 4: Visual-Need Router. If the answer is not confident, the Visual-Need Router evaluates whether pixel-level inspection of e^* would help, following eq. (5). If the router declines, $\mathcal{W}$ is cleared and the loop returns to Step 2 for the next iteration.

Step 5: On-demand frame extraction. When the router accepts, keyframes are adaptively sampled from e^*. The sampling rate is f_{ev}, capped at N_{fr} frames per event; when the frame count exceeds N_{fr}, frames are redistributed uniformly across the event duration. Extracted frames are pushed into the FIFO working memory $\mathcal{W}$ with capacity N_{wm}.

Step 6: Multimodal re-answer. The Answering Agent receives the enriched narrative memory (with all cumulative clip injections) together with the visual evidence in $\mathcal{W}$, in a single multimodal prompt. If confident, the answer is returned.

Fallback. If T iterations are exhausted without a confident answer, a must-answer fallback forces the agent to commit to its best answer using all accumulated textual and visual evidence.

Steps 2–6 repeat, and with each iteration the system's memory of the video becomes progressively richer: new events are explored (breadth), their clip details are permanently injected (depth), and visual evidence accumulates in $\mathcal{W}$ (grounding). The full procedure is summarized as pseudocode in Algorithm 1 of the appendix.

3 Experiment

We evaluate CFD on two long-video benchmarks: **Video-MME** (Fu et al., 2025) under the standard *without-subtitle* setting, and **InfiniBench** (Ataallah et al., 2025) on its four grounding-based skills (*Chronological Understanding*, *Character Actions Tracking*, *Scene Transitions*, and *Global Appearance*). Full dataset descriptions are provided in Appendix B.

3.1 Implementation details

We use `Qwen3-VL-8B-Instruct` (Bai et al., 2025a) as the frozen edge-side Captioning Agent (MLLM) and `Qwen3-VL-32B-Instruct` (Bai et al., 2025a) as the cloud-side reasoning MLLM that instantiates the Answering Agent, Localization Agent, and Visual-Need Router. Both models are served via vLLM (Kwon, 2025). **Offline indexing (edge).** For the event memory, we apply TransNetV2 (Soucek and Lokoc, 2024) (threshold 0.5) for shot-boundary detection, with a minimum segment merging threshold $\tau_{\min} = 60\,\text{s}$. Event-level captions use a structured prompt (scene overview, entity listing, chronological event flow, and anchor tags) at 1 FPS. For the clip memory, we use uniform clips of $\tau_c = 30\,\text{s}$ with stride equal to τ_c (non-overlapping), captioned at 1 FPS with a micro-action prompt (one action per sentence). Both memory tracks are cached under a deterministic configuration fingerprint and reused across queries. **Online reasoning (cloud).** The maximum number of backtracking iterations is $T = 5$. For on-demand frame extraction in Step 5, we use a sampling rate of $f_{\text{ev}} = 0.1$ FPS (1 frame per 10 s of event duration), capped at $N_{\text{fr}} = 8$ frames per event. When the ideal frame count exceeds N_{fr}, frames are redistributed uniformly across the event span. The FIFO visual working memory has capacity $N_{\text{wm}} = 16$. The Visual-Need Router uses a perceptual-versus-temporal rubric and operates as a single LLM call with structured JSON output. Detailed agent prompt designs are provided in Appendix F.

3.2 Comparison with other methods

Table 3 reports results on Video-MME. Open-sourced MLLMs typically feed hundreds to near-thousands of densely sampled frames; for instance, Qwen3-VL-32B reaches 75.9 overall at 768 frames per question. CFD achieves **67.5** overall with only **5.8 frames per question** on average, more than two orders of magnitude fewer than dense baselines. This asymmetry suggests that competitive temporal reasoning does not require saturating the context window with visual tokens, provided the right frames are retrieved at the right time. Among agent-based systems, VideoAgent (Fan et al., 2024), VideoTree (Wang et al., 2025d), DrVideo (Ma et al., 2025), and VideoLucy (Zuo et al., 2025) densely caption the full video during preprocessing and further revisit raw frames at query time to refine

Table 3: Performance comparison on Video-MME (Fu et al., 2025). † denotes reproduced results under strictly identical setups using the authors' code. *#Captions* denotes the average temporal span covered by each caption. *Re-caption?* indicates whether a method revisits video frames at query time to (re)caption or refine its textual memory. *#Frames* counts the number of frames fed to the answering model at the final inference stage.

Method	#Captions	Re-caption?	#Frames	Video-MME			
				short	medium	long	overall
Leading Open-source MLLMs							
VideoChat2-7B (Li et al., 2024b)	-	-	16	48.3	37.0	33.2	39.5
LongVA-7B (Zhang et al., 2024b)	-	-	128	61.1	50.4	46.2	52.6
Kangaroo-7B (Liu et al., 2024)	-	-	64	66.1	55.3	46.6	56.0
Video-CCAM-14B (Fei et al., 2024)	-	-	96	62.2	50.6	46.7	53.2
VideoXL-7B (Shu et al., 2025)	-	-	128	64.0	53.2	49.2	55.5
Dispider-7B (Qian et al., 2025)	-	-	1 fps	-	-	-	57.2
VideoChat-Online-4B (Huang et al., 2025)	-	-	2 fps	-	-	47.1	54.4
TimeChat-Online-7B (Yao et al., 2025)	-	-	1 fps	-	-	48.4	62.4
Qwen3-VL-32B (Bai et al., 2025a)	-	-	768	82.4	76.3	69.0	75.9
Agent-based Systems							
VideoAgent (Fan et al., 2024)	2s / caption	✓	-	-	-	46.4	-
VideoTree (Wang et al., 2025d)	8s / caption	✓	-	67.8	59.9	54.2	60.6
DrVideo (Ma et al., 2025)	5s / caption	✓	-	-	-	51.7	-
MemVid (Yuan et al., 2025)	-	✗	1 fps	**73.9**	63.1	55.0	64.0
VideoLucy (Zuo et al., 2025)†	< 30s / caption	✓	-	73.2	64.7	56.2	64.7
CFD (Ours)	**233.1s / caption**	✗	**5.8**	72.2	**66.6**	**63.6**	**67.5**

Table 4: Performance comparison on InfiniBench (Ataallah et al., 2025). *#Captions* denotes the average temporal span covered by each caption. *#Frames* counts the number of frames fed to the answering model at the final inference stage. *Chronological Understanding* evaluates ordering events across the full video; *Character Actions Tracking* requires aggregating and sequencing a character's actions; *Scene Transitions* tests recognizing and ordering scene/location shifts; *Global Appearance* tracks long-term appearance changes of a character (e.g., outfit changes).

Method	#Captions	#Frames	Chronological Understanding	Global Appearance	Scene Transitions	Character Actions
LLaVA-OneVision (Li et al., 2024a)	-	128	43.91	37.21	25.40	20.11
InternVL2.5 (Zhu et al., 2025)	-	128	42.16	34.88	20.63	21.60
Qwen2-VL (Wang et al., 2024a)	-	768	48.41	31.01	23.81	32.59
Qwen2.5-VL (Bai et al., 2025b)	-	768	27.17	37.98	22.22	22.53
InternVL3 (Zhu et al., 2025)	-	128	36.29	33.33	28.57	23.46
Qwen3-VL (Bai et al., 2025a)	-	768	48.44	70.54	53.97	67.04
CFD (Ours)	206s / caption	14.8	55.10	58.90	52.40	56.20

their textual memory, while MemVid (Yuan et al., 2025) avoids re-captioning but compensates with dense 1 fps uniform sampling (64.0 overall, 73.9 on short videos). CFD matches or exceeds all reported agent-based baselines (VideoLucy 64.7, MemVid 64.0) while using roughly an order of magnitude fewer frames per question. We do not aim to match dense-inference upper bounds such as Qwen3-VL-32B at 768 frames; we target the accuracy-efficiency frontier under bounded visual budgets. The gains over agent-based methods are most pronounced on medium and long videos (66.6 and 63.6), where re-captioning and dense sampling suffer most from context saturation and redundant visual processing.

Table 4 reports results on the four grounding-based skills of InfiniBench. CFD achieves the best Chronological Understanding score (55.10), exceeding the 768-frame dense baselines Qwen3-VL (48.44) and Qwen2-VL (48.41), and remains competitive on Scene Transitions (52.40 *vs*. 53.97 for Qwen3-VL), at only **14.8 frames per question** on average. These two skills are most aligned with our design principle: ordering events and recognising scene-level shifts are temporally structured tasks for which the event-level narrative skeleton provides a compact discrete record. On Global Appearance and Character Actions, CFD (58.90 / 56.20) is ahead of agent-based baselines but below the dense Qwen3-VL, consistent with the visual-

Table 5: Component Study. (*) denotes that all clip captions in one video are used for the ablation experiment.

Event	Clip	ReasoningLoop	Frames	Router	Video-MME				InfiniBench			
					short	medium	long	overall	Chronological Understanding	Global Appearance	Scene Transitions	Character Actions
✓					62.0	53.7	56.3	57.3	52.30	9.30	39.70	16.90
	✓*				68.2	66.4	62.0	65.5	55.70	38.00	65.10	43.90
✓	✓	✓			68.8	64.0	63.0	65.3	55.10	45.00	52.40	53.40
✓	✓	✓	✓		72.3	66.7	62.6	67.2	51.70	62.80	58.70	65.90
✓	✓	✓	✓	✓	72.2	66.6	63.6	67.5	55.10	58.90	52.40	56.20

Table 6: Comparison across different captioner model sizes.

Caption Model	Size	short	medium	long	overall
Qwen3-VL	2B	67.7	61.9	58.1	62.6
	8B	72.2	66.6	63.6	67.5
	32B	72.4	67.1	61.0	66.8

Table 7: Comparison across different budget settings.

N_{fr}	N_{wm}	short	medium	long	overall
2	4	71.8	66.3	63.6	67.2
8	16	72.2	66.6	63.6	67.5
32	64	72.1	66.6	64.2	67.6

textual duality identified in Section 1: fine-grained attribute binding benefits from dense pixel access, and a tight on-demand budget does not fully substitute for exhaustive sampling. Even so, within the agent-based regime the Visual-Need Router still recovers a substantial fraction of the appearance-grounding evidence under a tight frame budget.

3.3 Ablation study

Component Analysis. Table 5 ablates each component. Event memory alone is too sparse (57.3 overall, Global Appearance 9.30). Dense clip memory alone (Row 2) recovers most of the gap (65.5 overall, Chronological Understanding 55.70) but Global Appearance remains limited (38.00), confirming that text alone cannot substitute for visual evidence. Adding the reasoning loop over the dual-track memory without any frames (Row 3) achieves comparable overall performance (65.3), corroborating that language memories are a strong carrier for temporal structure while attribute perception remains the bottleneck of text-only reasoning. Enabling on-demand frames without the router (Row 4) brings sharp gains on appearance-centric tasks (Global Appearance 62.80, Character Actions 65.90) at the cost of Chronological Understanding (51.70) and long-video accuracy (62.6), reflecting that indiscriminate frame injection dilutes the narrative context on temporally structured questions. The full model with the Visual-Need Router (Row 5) suppresses frames on those questions, recovering Chronological Understanding to 55.10 and long-video accuracy to 63.6 while keeping the average frame count low, achieving a more balanced accuracy-efficiency trade-off than always-on frame retrieval. This confirms that text and frames are complementary: text excels at long-horizon temporal structure, while frames remain decisive for fine-grained attribute perception, and routing between them captures the best of both regimes.

Offline captioner model size. Table 6 sweeps three Qwen3-VL captioner sizes used during offline indexing. Scaling `2B`→`8B` brings a substantial gain (62.6→67.5 overall); scaling further to `32B` marginally helps short and medium videos but degrades long-video accuracy (61.0 *vs.* 63.6). We attribute this to verbosity accumulation: larger captioners produce longer, more detailed descriptions per segment, which compound across the many events of a long video and inflate the total narrative context beyond what the reasoning agent can attend to within a fixed inference budget. We therefore adopt `Qwen3-VL-8B` as the default, balancing indexing cost, caption conciseness, and downstream accuracy. The `Qwen3-VL-2B` captioner reaches 62.6 overall and provides a lower-resource alternative for more constrained offline indexing tiers.

Budget sensitivity. Table 7 sweeps the per-event frame cap N_{fr} and FIFO capacity N_{wm} over an 8× range. Overall accuracy varies by only 0.4%, indicating that a few targeted frames suffice; the middle setting (N_{fr}=8, N_{wm}=16) is within 0.1% of the largest budget at one-quarter the visual cost. Long videos benefit most (63.6→64.2), consistent with sparser event memory leaving more room for visual verification. These results support the core premise of CFD: a small, bounded set of on-demand frames closes most of the gap that narrative memory alone cannot cover.

Frame reduction and routing. Even when the Router is disabled and every localized event is routed to frames, the caption-once memory, tem-

Table 8: Visual-budget sweep on Video-MME.

N_{fr}	N_{wm}	Frames/query	Accuracy
2	4	1.83	67.2
8	16	5.76	67.5
32	64	8.93	67.6

Table 9: Controlled routing comparison on Video-MME. Each cell is (accuracy / online frames per query).

Policy	Short	Medium	Long	Average
No frames	68.8 / 0.00	64.0 / 0.00	63.0 / 0.00	65.3 / 0.00
Visual-Need Router	72.2 / 2.38	66.6 / 5.81	63.6 / 9.09	67.5 / 5.76
Always frames	72.3 / 2.54	66.7 / 6.57	62.6 / 10.78	67.2 / 6.63

poral localization, iterative backtracking, bounded frame extraction, and fixed-capacity visual working memory reduce online frame usage from 768 to 6.63 frames/query, corresponding to **99.1% fewer frames**, or approximately a **116**× reduction. Within this already sparse pipeline, the Visual-Need Router further reduces frame usage from 6.63 to 5.76 frames/query, a further **13.1% reduction**, while improving accuracy from 67.2 to 67.5 (**+0.3 percentage points**). Frame savings do not imply lower wall-clock latency in every configuration, because the prompt-based Router introduces an additional LLM call. The matched budget sweep reports the corresponding online frame counts in Table 8. Increasing the budget from 2/4 to 8/16 improves accuracy by 0.3 points, while further increasing it to 32/64 yields only another 0.1 point despite a substantially larger frame footprint. We therefore treat 8/16 as an accuracy–cost operating point rather than exhaustive visual coverage.

Routing policy. Table 9 reports the controlled comparison among no-frame, Visual-Need Router, and always-frame inference. The Router improves average accuracy by **2.2 points** over no-frame reasoning and slightly exceeds always-frame inference (**67.5 vs. 67.2**) while reducing online frames by **13.1%** on average. The full policy table and outcome-oracle analysis are provided in Appendix E. Table 16 in Appendix E reports trigger statistics from the same paired diagnostic rerun used to construct the outcome oracle. The Router retrieves frames for **46.2%** of questions and therefore leaves **53.8%** entirely in language space, while capturing 127 of 133 (**95.5%**) cases in which frames empirically improve the answer. These cases are outcome-based diagnostics, not human routing annotations.

Iteration budget. Figure 3 shows performance as a function of the maximum iteration budget T. Short-video accuracy saturates rapidly, reaching 71.9 at

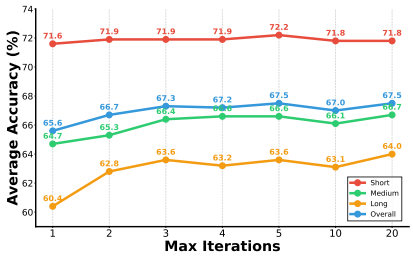

Figure 3: Effect of the maximum iteration budget T on Video-MME accuracy across short, medium, long, and overall splits.

Table 10: Clip-caption activation on Video-MME. Counts are captions per query, not seconds per caption or captioning runtime.

Method	Short	Medium	Long	Average
All 30-s clip captions	3.10	17.69	82.66	34.49
CFD-activated captions	1.52	9.29	24.05	11.62
Caption reduction	51.2%	47.5%	70.9%	66.3%

T=2, while medium-video performance improves until T=4 (66.6) and then plateaus. Long-video accuracy continues improving from 60.4 at T=1 to 64.0 at T=20, indicating that longer videos benefit from additional retrieval iterations. We adopt T=5 as the default, which achieves the best overall score of 67.5 and provides a favorable trade-off across duration splits.

Clip-caption activation. As shown in Table 10, CFD uses **66.3% fewer clip captions on average** while outperforming the all-clip baseline overall. Event Memory uses TransNetV2 boundaries, whereas Clip Memory is generated uniformly and independently of event segmentation. The outcome oracle uses frames only when the always-frame policy corrects an error made by the no-frame policy; it is an empirical upper bound rather than semantic ground-truth annotation. Its 133 frame-beneficial cases correspond to a **4.9%** trigger rate over all 2,700 questions, distinct from the Router's 46.2% YES rate. The oracle obtains 69.5% accuracy with 0.64 frames/query, compared with the Router's 67.5 accuracy and 5.76 frames/query.

4 Conclusion

We presented CFD, a budget-aware edge-cloud framework that indexes video into reusable narrative memory once and retrieves sparse visual frames only when language alone cannot ground the answer. Experiments across long-video benchmarks confirm that this caption-once, frames-on-demand design achieves competitive accuracy while substantially reducing online visual cost.

Limitations

Caption quality and segmentation sensitivity. Our framework is fundamentally bottlenecked by the quality of the offline captioning pass. If question-critical visual details are not captured in the question-agnostic memories, purely text-based backtracking cannot recover them afterwards. Sparse visual backtracking partially alleviates this issue, but its effectiveness still depends on whether the system can correctly diagnose why text-only reasoning failed and route the query to the appropriate event. Moreover, even after the correct event is localized, the bounded keyframe extraction stage may still miss highly transient cues, small objects, or brief textual overlays that fall outside the sampled frames. It may also miss subtle motion. Failure of both textual localization and sparse visual retrieval therefore remains possible. The framework is also sensitive to the quality of upstream event segmentation: overly coarse segmentation can bury useful evidence inside broad summaries, whereas overly fine segmentation may fragment semantically coherent episodes and weaken later localization. On the modeling side, both the captioner and the Visual-Need Router currently operate in a generic zero-shot setting, and would likely benefit from domain-specific supervision or lightweight fine-tuning to improve structured caption fidelity and routing accuracy. Finally, our edge-cloud setup is validated under simulated separation rather than physical deployment, so the practical impact of device-side compute constraints, communication overhead, and network latency variability remains to be systematically characterized on representative edge hardware.

Router granularity. The current Visual-Need Router operates as a binary, zero-shot classifier over a perceptual-versus-temporal rubric. While this coarse dichotomy is sufficient to demonstrate that question-type-conditioned visual gating yields meaningful cost savings, it inevitably mis-routes queries that fall on the boundary (e.g., a chronological question whose answer hinges on a specific visual change). A continuous, calibrated confidence score, possibly trained from a small held-out routing set or distilled from oracle traces, would likely improve per-skill balance, particularly on InfiniBench's attribute-heavy skills where the current router under-allocates frames relative to the no-router configuration (Row 4 of Table 5).

Egocentric and content-dependent segmentation. Our event memory relies on TransNetV2 shot-boundary detection, which is effective for edited video (TV shows, documentaries, lectures) but is not designed for continuous egocentric or wearable recordings where hard cuts are rare. Adapting CFD to egocentric streams would require replacing shot-boundary segmentation with an audio-visual event detector (e.g., EgoTrigger (Paruchuri et al., 2025)) or an LLM-prompted scene-change estimator over coarse-frame thumbnails; both are orthogonal to our framework and represent natural follow-up work.

Definition of "edge". We use "edge" to denote a separation of inference tiers (offline indexing vs. online cloud reasoning) rather than a claim that the captioner runs on mobile-class hardware. Our default captioner (Qwen3-VL-8B) targets a modest server-class device or a powerful desktop with a single GPU; deploying onto smartphones or smart glasses would require additional model compression (e.g., Qwen3-VL-2B with quantisation), which we leave to future work alongside end-to-end latency and energy measurements on representative hardware.

References

Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. 2025. Infinibench: A benchmark for large multi-modal models in long-form movies and tv shows. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19485–19512.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Weitong Cai, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. 2025. Mllm as video narrator: Mitigating modality imbalance in video moment retrieval. *Pattern Recognition*, 166:111670.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024a. Timemarker: A versatile video-

llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Hyomin Choi and Ivan V Bajić. 2018. Deep feature compression for collaborative object detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3743–3747. IEEE.

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 *ECCV*, pages 75–92. Springer.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. 2024. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*.

Chaoyou Fu, Yuhan 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 *CVPR*, pages 24108–24118.

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

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. 2025. Online video understanding: Ovbench and videochat-online. In *CVPR*, pages 3328–3338.

Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, et al. 2025. Storm: Token-efficient long video understanding for multimodal llms. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5889–5900. IEEE.

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.

Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News*, 45(1):615–629.

Woosuk Kwon. 2025. *vLLM: An Efficient Inference Engine for Large Language Models*. Ph.D. thesis, UC Berkeley.

Pingrui Lai, Zihao Xie, and Hua Yang. 2025. Haven: From human guidance to assistant by evolution network in vision-and-language navigation. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206.

Wei Li, Hehe Fan, Yongkang Wong, Mohan Kankanhalli, and Yi Yang. 2024c. Topa: Extending large language models for video understanding via text-only pre-alignment. *NeurIPS*, 37:5697–5738.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. 2026. Videochat-flash: Hierarchical compression for long-context video modeling. In *ICLR*, volume 2026, pages 109089–109117.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. 2025. Drvideo: Document retrieval based long video understanding. In *CVPR*, pages 18936–18946.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 36:46212–46244.

Akshay Paruchuri, Sinan Hersek, Lavisha Aggarwal, Qiao Yang, Xin Liu, Achin Kulshrestha, Andrea Colaco, Henry Fuchs, and Ishan Chatterjee. 2025. Egotrigger: Toward audio-driven image capture for human memory enhancement in all-day energy-efficient smart glasses. *IEEE Transactions on Visualization and Computer Graphics*.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *CVPR*, pages 24045–24055.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024. Streaming long video understanding with large language models. *NeurIPS*, 37:119336–119360.

Qwen Team. 2025. Qwen3.5. https://qwen.ai/blog?id=qwen3.5.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. 2026. Agentic very long video understanding. *arXiv preprint arXiv:2601.18157*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *CVPR*, pages 26160–26169.

Raghubir Singh and Sukhpal Singh Gill. 2023. Edge ai: a survey. *Internet of Things and Cyber-Physical Systems*, 3:71–92.

Tomás Soucek and Jakub Lokoc. 2024. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221.

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

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. 2025a. Lvbench: An extreme long video understanding benchmark. In *ICCV*, pages 22958–22967.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. 2024b. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*.

Xiao Wang, Qingyi Si, Shiyu Zhu, Jianlong Wu, Li Cao, and Liqiang Nie. 2025c. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5417–5432.

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.

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. 2025d. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *CVPR*, pages 3272–3283.

Siwei Wen, Zhangcheng Wang, Xingjian Zhang, Lei Huang, and Wenjun Wu. 2026. Eventmemagent: Hierarchical event-centric memory for online video understanding with adaptive tool use. *arXiv preprint arXiv:2602.15329*.

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

Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. 2025a. Captionqa: Is your caption as useful as the image itself? *arXiv preprint arXiv:2511.21025*.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2025b. Vca: Video curious agent for long video understanding. In *ICCV*, pages 20168–20179.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. 2025. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10807–10816.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. 2025. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. *arXiv preprint arXiv:2512.12360*.

Huaying Yuan, Zheng Liu, Minghao Qin, Hongjin Qian, Yan Shu, Zhicheng Dou, Ji-Rong Wen, and Nicu Sebe. 2025. Memory-enhanced retrieval augmentation for long video understanding. *arXiv preprint arXiv:2503.09149*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. 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.

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

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. 2025b. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*.

Zhuo Zhi, Qiangqiang Wu, Minghe shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. 2025. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. *arXiv preprint arXiv:2504.04471*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. 2025. Mlvu: Benchmarking multi-task long video understanding. In *CVPR*, pages 13691–13701.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. 2025. Videolucy: Deep memory backtracking for long video understanding. *arXiv preprint arXiv:2510.12422*.

This appendix complements the main paper with additional analyses, organized as follows.

Appendix A (Reasoning Loop Algorithm) presents the pseudocode of the story-first reasoning loop with on-demand frame retrieval that orchestrates the four agent roles described in the main paper.

Appendix B (Datasets) provides full descriptions of the Video-MME and InfiniBench benchmarks used in the main experiments, together with the evaluation protocol and the rationale for focusing on InfiniBench's grounding-based skills.

Appendix C (Related Work) positions CFD with respect to prior work on long video understanding with MLLMs, agentic video understanding, and video understanding on edge devices.

Appendix D (Results on LVBench) extends the main evaluation to LVBench, a challenging hour-scale benchmark, demonstrating the generalizability of CFD.

Appendix E (Further Ablation Study) studies cloud reasoning model size, cross-family backbone generalization, component and budget sensitivity, inference efficiency, practical cost accounting, and Visual-Need Router reliability. It also presents qualitative visualizations of representative inference trajectories illustrating the story-first, evidence-on-demand behavior.

Appendix F (More Implementation Details) supplies the full agent prompts used in our system.

Appendix G (Licenses of Datasets and Models) documents the licenses of all benchmarks, pre-trained models, and software artifacts used in this work, and confirms that our usage is consistent with their intended scope.

Appendix H (Use of AI Assistants) discloses the scope of AI-assistant usage during the preparation of this paper.

A Reasoning Loop Algorithm

For completeness, Algorithm 1 presents the pseudocode of the story-first reasoning loop with on-demand frame retrieval described in Section 2.3 of the main paper. The procedure operates over the dual-track narrative memory $(\mathcal{M}_E, \mathcal{M}_C)$ and the visual working memory $\mathcal{W}$, invoking the Answering, Localization, and Visual-Need Router agents under a bounded iteration budget T and a fixed frame budget governed by N_{fr} and the FIFO capacity N_{wm}.

B Datasets

Video-MME. Video-MME (Fu et al., 2025) is a comprehensive multimodal evaluation benchmark for video understanding, containing 900 manually curated videos and 2,700 expert-annotated multiple-choice questions (3 per video). The benchmark covers 6 primary visual domains with 30 fine-grained subcategories and spans a wide range of durations: short (<2 min, avg. 82.5s), medium (4–15 min, avg. 562.7s), and long (30–60 min, avg. 2,385.5s). Question types encompass temporal perception, spatial reasoning, action recognition, object recognition, and information synopsis, among others. All experiments follow the standard *without-subtitle* setting, and we report accuracy on each duration split as well as the overall average.

InfiniBench. InfiniBench (Ataallah et al., 2025) is a large-scale benchmark targeting long-form video understanding in movies and TV episodes, featuring an average video duration of 53 minutes and over 87.7K question-answer pairs. We evaluate our system exclusively on its four *grounding-based* skills, which assess a model's ability to retrieve, order, and structure video content without requiring causal inference: *Chronological Understanding* measures the ability to correctly sequence a set of events across the full video. *Character Actions Tracking* requires grouping and ordering all actions performed by a specific character over the video duration. *Scene Transitions* tests recognition and sequential ordering of scene-level location shifts. *Global Appearance* evaluates long-term tracking of changes in a character's visual appearance (*e.g.*, the sequence of outfit changes throughout the video). All four skills are posed as multiple-choice questions and evaluated with standard classification accuracy. We focus on these skills because they directly probe temporally-extended visual localization, the core capability our system is designed to address, without conflating performance with the higher-level causal and narrative inference required by the reasoning-based skills.

C Related Work

Long video understanding with MLLMs. Benchmarks from EgoSchema (Mangalam et al., 2023) to hour-scale suites such as Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), LVBench (Wang et al., 2025a), and LongVideoBench (Wu et al., 2024) consistently show that performance degrades with duration, motivating work along two axes. *Vi-*

Algorithm 1 Story-First Reasoning with On-Demand Frames

Require: Video V, question Q, memories $\mathcal{M}_E$, $\mathcal{M}_C$, agents ANSAGENT, LOCAGENT, ROUTER, budget T, FIFO capacity N_{wm}, max frames N_{fr}, sampling rate f_{ev}.

$\mathcal{W} \leftarrow \varnothing$; $\mathcal{S}_{\text{tried}} \leftarrow \varnothing$; $\mathcal{B} \leftarrow \varnothing$ ▷ working memory, explored set, clip bank

▷ **Step 1**: Story-first answer

$(a, \mathit{conf}, r) \leftarrow \text{ANSAGENT}(\mathcal{M}_E, \mathcal{W}, Q)$

if *conf* **then return** a

end if

for $i = 1$ **to** T **do**

▷ **Step 2**: Temporal localization

$e^* \leftarrow \text{LOCAGENT}(\mathcal{M}_E \setminus \mathcal{S}_{\text{tried}},\ \mathcal{W},\ Q)$

$\mathcal{S}_{\text{tried}} \leftarrow \mathcal{S}_{\text{tried}} \cup \{e^*\}$

▷ **Step 3**: Clip-enriched re-answer

$\mathcal{B} \leftarrow \mathcal{B} \cup \{(e^*,\ \{m_j^C : c_j \cap e^* \neq \varnothing\})\}$ ▷ accumulate clip enrichments

$\mathcal{M} \leftarrow \text{REBUILD}(\mathcal{M}_E,\ \mathcal{B})$ ▷ inject all accumulated clips

$(a, \mathit{conf}, r) \leftarrow \text{ANSAGENT}(\mathcal{M},\ \mathcal{W},\ Q)$

if *conf* **then return** a

end if

▷ **Step 4**: Visual-Need routing

need_visual $\leftarrow \text{ROUTER}(e^*,\ Q,\ r)$

if $\neg$ *need_visual* **then**

$\mathcal{W} \leftarrow \varnothing$; **continue** ▷ skip frames, next iteration

end if

▷ **Step 5**: On-demand frame extraction

$\mathcal{F} \leftarrow \text{ADAPTIVESAMPLE}(V,\ e^*,\ f_{\text{ev}},\ N_{\text{fr}})$

$\mathcal{W} \leftarrow \text{FIFO_PUSH}(\mathcal{W},\ \mathcal{F},\ N_{\text{wm}})$

▷ **Step 6**: Multimodal re-answer

$(a, \mathit{conf}, r) \leftarrow \text{ANSAGENT}(\mathcal{M},\ \mathcal{W},\ Q)$

if *conf* **then return** a

end if

end for

return $\text{MUSTANSWER}(\mathcal{M},\ \mathcal{W},\ Q)$ ▷ Fallback

sual compression methods reduce token count so that more frames fit within an MLLM's context: Video-XL (Shu et al., 2025) condenses KV states across intervals, LongVU (Shen et al., 2024) filters redundant frames via DINOv2 similarity, and VideoChat-Flash (Li et al., 2026) applies hierarchical token merging (Bolya et al., 2022). *Textual translation* methods instead convert video into language surrogates: LLoVi (Zhang et al., 2024a) chains dense captioning with LLM summarization, LangRepo (Kahatapitiya et al., 2025) maintains a prunable language repository, and TOPA (Li et al., 2024c) demonstrates temporal reasoning from synthetic textual videos alone. Text compactly encodes temporal topology but systematically loses fine-grained visual attributes (Yang et al., 2025a), motivating our factorization of language narrative and on-demand visual evidence. The observation that captions can carry much of the long-range temporal reasoning load is already implicit in LLoVi (EgoSchema results) and made explicit in TOPA; we therefore treat this visual-textual duality as established prior context rather than a primary contribution, and focus instead on how to systematically *gate* visual access via a query-conditioned router so that the duality is exploited under explicit cost control rather than left to emergent pipeline behaviour.

Agentic video understanding. Rather than processing all frames at once, agent-based systems treat videos as searchable environments. Early examples include VideoAgent (Fan et al., 2024) (iterative planning and retrieval), DrVideo (Ma et al., 2025) (document-augmented agent loops), and VideoTree (Wang et al., 2025d) (query-adaptive keyframe trees). Recent work introduces richer memory and search: VideoLucy (Zuo et al.,

2025) proposes hierarchical memory with iterative backtracking; VideoARM (Yin et al., 2025), DVD (Zhang et al., 2025b), HAVEN (Lai et al., 2025), EGAgent (Rege et al., 2026), and EventMemAgent (Wen et al., 2026) further explore dynamic multimodal memory, multi-granular toolsets, entity-aware indexing, and online memory with agentic RL. Ours shares the evidence-gathering paradigm but *factorises* the representation into a reusable narrative index plus on-demand visual evidence, and explicitly optimizes for edge-cloud cost rather than accuracy alone. Several recent agentic systems (DVD (Zhang et al., 2025b), EGAgent (Rege et al., 2026), EventMemAgent (Wen et al., 2026)) appeared concurrently with this work; we do not include head-to-head numbers because their public implementations were not yet available at the time of submission, and we leave systematic cross-comparison to follow-up work.

Video understanding on edge devices. Collaborative intelligence (Kang et al., 2017) and deep feature compression (Choi and Bajić, 2018) study DNN partitioning and intermediate-feature codecs for split inference but target single-inference offloading. EgoTrigger (Paruchuri et al., 2025) gates camera activation on smart glasses via audio cues, reducing capture by 54% while preserving QA accuracy. Our framework operates at a higher semantic level: instead of gating the camera, we gate the *representation* sent to the cloud, transmitting compact narrative memory by default and reserving frame uploads for moments requiring visual attribution.

D Results on LVBench

In this section, we conduct more comprehensive comparisons on the LVBench (Wang et al., 2025a). LVBench (Wang et al., 2025a) is a benchmark specifically designed for long video understanding. It contains 103 manually curated long-form YouTube videos spanning 117 hours in total, with an average duration of 4,101 seconds per video, making it substantially longer than most existing video understanding benchmarks. The videos cover six diverse domains, including sports, documentary, event record, lifestyle, TV shows, and cartoons, and are paired with 1,549 human-annotated question-answer instances. To comprehensively evaluate long-range video comprehension, LVBench organizes evaluation around six core capabilities: entity recognition, event understanding, key information retrieval, temporal grounding, reasoning, and summarization. An additional strength of LVBench is its high-quality annotation protocol, where questions are manually designed to require visual evidence and are accompanied by temporal clues indicating the minimal relevant video segment. These properties make LVBench a challenging and representative benchmark for assessing multimodal models under extended temporal contexts.

Table 11 presents results on LVBench (Wang et al., 2025a). CFD achieves an overall score of 52.9, outperforming all agent-based systems that use open-source models, including MemVid (44.4), VCA (41.3), VideoTree (28.8), and VideoAgent (29.3), and remaining competitive with the leading open-source MLLM AdaReTaKe-72B (53.3) despite operating through a bounded agentic pipeline rather than dense end-to-end inference. VideoLucy (58.8) employs DeepSeek-R1 (Guo et al., 2025) as its agent base model (denoted †), yet CFD narrows the gap to 5.9 points using only open-source components throughout, demonstrating that our caption-once, frames-on-demand design remains competitive even against pipelines backed by frontier closed-source reasoning models. Among the six evaluation dimensions, CFD achieves its strongest result on Key Information Retrieval (63.9), where the dual-track narrative index provides precisely timestamped textual anchors that facilitate accurate localization of query-relevant content.

Notably, CFD uses identical hyperparameters across all evaluated benchmarks without any dataset-specific tuning, indicating that the caption-once dual-track indexing and the visual-need routing strategy generalize robustly across substantially different video durations and evaluation protocols, from the short-to-long splits of Video-MME to the hour-long narratives of LVBench and InfiniBench.

E Further Ablation Study

Cloud reasoning model size. Table 12 ablates the size of the cloud-side MLLM responsible for the Answering Agent, Localization Agent, and Visual-Need Router. Performance scales consistently with model size across all duration splits, from 49.7 overall at 2B to 67.5 at 32B, confirming that reasoning quality and routing accuracy are meaningful bottlenecks in the pipeline. The gap between 2B and 8B (9.2 points overall) is substantially larger than that between 8B and 32B (8.6 points), suggesting diminishing returns at the upper end. In contrast to the

Table 11: Performance comparison on LVBench (Wang et al., 2025a) across six evaluation dimensions: *Entity Recognition* (ER), *Event Understanding* (EU), *Key Information Retrieval* (KIR), *Temporal Grounding* (TG), *Reasoning* (Rea), and *Summarization* (Sum). *Overall* denotes the average score over all six dimensions. † denotes methods that employ a closed-source LLM as the agent base model.

Method	ER	EU	KIR	TG	Rea	Sum	Overall
Leading Open-source MLLMs							
TimeMarker-8B (Chen et al., 2024a)	42.8	39.1	34.9	38.7	38.2	48.8	41.3
VideoLLaMA3-7B (Zhang et al., 2025a)	45.8	42.4	47.8	35.9	45.8	36.2	45.3
InternVL2.5-78B (Chen et al., 2024b)	43.8	42.0	42.1	36.8	51.0	37.9	43.6
Qwen2-VL-72B (Wang et al., 2024a)	38.0	41.1	38.3	41.4	46.5	46.6	41.3
ReTake-7B (Wang et al., 2024b)	49.8	46.2	52.9	45.0	45.8	27.6	47.8
VideoChat-Flash-7B (Li et al., 2026)	51.1	46.0	49.0	38.9	48.5	34.5	48.2
AdaReTaKe-72B (Wang et al., 2025c)	53.0	50.7	62.2	45.5	54.7	37.9	53.3
InternVL3-8B (Zhu et al., 2025)	47.7	43.0	42.6	42.3	46.8	25.9	44.5
Agent-based Systems							
VideoAgent (Wang et al., 2024c)	28.0	30.3	28.0	29.3	28.0	36.4	29.3
VideoTree (Wang et al., 2025d)	30.3	25.1	26.5	27.7	31.9	25.5	28.8
MemVid (Yuan et al., 2025)	53.4	40.6	46.3	34.9	43.2	28.1	44.4
VCA (Yang et al., 2025b)	43.7	40.7	37.8	38.0	46.2	27.3	41.3
VideoLucy (Zuo et al., 2025)†	54.3	59.8	75.6	51.7	55.9	49.1	58.8
CFD (Ours)	53.9	50.2	63.9	48.2	48.3	37.9	52.9

Table 12: Video-MME accuracy with different cloud reasoning model sizes.

Reasoning Model	Size	short	medium	long	overall
Qwen3-VL	2B	57.3	47.7	44.1	49.7
	8B	61.6	61.6	53.6	58.9
	32B	72.2	66.6	63.6	67.5

captioner size ablation (Table 6 in the main paper), where scaling to 32B slightly degraded long video performance due to caption verbosity accumulation, the cloud reasoning model benefits monotonically from scale, as larger models provide more reliable event localization and more accurate routing decisions without inflating the narrative context. We further observe that the 2B model frequently terminates at Step 1 by returning a confident answer directly from the event-level story skeleton, bypassing the localization and visual verification stages entirely. This suggests that smaller models lack the calibration to assess their own uncertainty reliably, and that effective agentic workflows with iterative backtracking currently require sufficiently large reasoning models to engage the full pipeline.

Generalization across model families. To assess the generalizability of CFD beyond a single model family, we evaluate two additional backbone configurations: InternVL3.5 (Wang et al., 2025b), a widely adopted open-source MLLM family, and Qwen3.5 (Qwen Team, 2025), a recently released series with stronger reasoning capabilities. Table 13 shows that CFD with Qwen3-VL (8B captioner, 32B reasoner) achieves 67.5 overall at 5.8 frames per question, while the InternVL3.5 configuration (8B + 38B) reaches 64.3 at 7.6 frames, demonstrating that the pipeline transfers across model families without architectural changes. Replacing both the captioner and reasoner with Qwen3.5 (9B + 27B) further improves overall performance to **69.4** with only **2.5 frames** per question, outperforming all agent-based baselines and substantially closing the gap with Qwen3-VL-32B (75.9) at 768 frames. The lower frame count indicates that the Qwen3.5 backbone configuration triggers visual verification less frequently. Because both the captioner and reasoner are changed simultaneously, we do not attribute this reduction solely to routing quality; it may also reflect stronger textual memories, earlier confidence-based exits, or different localization behavior. Overall, these results demonstrate that CFD transfers across model families without architectural changes. Notably, Qwen3.5 is a natively multimodal model that processes interleaved text and visual inputs within a unified architecture, confirming that supplying a small number of on-demand frames alongside rich textual memories at reasoning time is both compat-

ible with and well-suited to the emerging paradigm of native multimodal reasoning.

Reasoning efficiency. Table 14 reports online LLM serving time for no-frame, always-frame, and routed inference. It is worth noting that the Router adds an LLM call, so frame savings do not imply lower wall-clock latency in every configuration.

Qualitative visualizations. Figure 4 presents the complementary *direct-answer* mode of CFD, where the question can already be resolved from coarse event memory without any further backtracking. In this example, the retrieved event memory explicitly summarizes the video as a behind-the-scenes record of a live theatre broadcast, covering crew operations, rehearsals, makeup, control-room monitoring, and audience viewing, while the prominent text repeatedly anchors the scene to *National Theatre Live*. As a result, the model answers correctly at the initial *story-first* stage with high confidence, selecting option **B** directly from the global semantic storyline. Unlike the first case, this question does not depend on a fine-grained perceptual attribute or a narrowly localized visual cue; instead, it is fully supported by high-level narrative evidence already available in the event memory. This example therefore demonstrates the efficiency-oriented side of CFD: when the memory abstraction is already sufficiently informative, the system stops early and avoids unnecessary event localization, routing, and visual evidence retrieval.

Figure 5 visualizes the full inference trajectory of CFD on a representative *attribute perception* example, where the question asks for the main color of *Nuna*. The case clearly illustrates the core design of our framework: *story first, evidence on demand*. Starting from coarse *Event Memory*, the model first attempts a story-level answer but correctly abstains, since the retrieved memory only indicates that "Nuna Supercontinent Forms" appears in the relevant event, while no explicit color cue is provided. It then performs temporal localization and selects the large event spanning [0.0, 505.0], after which a clip-enriched re-answer further narrows the relevant evidence to the local spans around [240,300] seconds, where *NUNA* explicitly appears, yet still remains uncertain because the textual memory does not fully specify the target attribute. At this point, the Visual-Need Router correctly identifies the question as a *local perceptual query*, rather than a long-horizon temporal reasoning problem, and triggers sparse evidence extraction only when textual memory is insufficient. The final multimodal re-answer then resolves the ambiguity by incorporating a small set of visual keyframes, leading to the correct prediction.

Together, the two examples highlight two important properties of CFD: first, it avoids unnecessary visual access when high-level memory is already sufficient; second, when the missing information is fine-grained and perceptual, it can selectively backtrack to visual evidence instead of reprocessing the entire video. Overall, the visualization demonstrates how our method progressively refines its evidence source from coarse narrative memory, to localized clip memory, and finally to sparse visual grounding, which is precisely the behavior needed for efficient long-video question answering.

Practical motivation and cost accounting. A representative application is interactive wearable lifelogging. Commercial products such as Looki L1 capture daily experiences hands-free and transform them into lifelogs, stories, videos, and searchable memories. In our *single-device, single-run hands-on test*, approximately 30 minutes of footage required around two hours to transfer from the device to the phone and then to the cloud, followed by more than five hours of preprocessing before AI interaction became available. We report this only as an illustrative observation under our specific device and network conditions, rather than as a general benchmark of the product. These observations motivate a clear separation between one-time indexing and interactive QA: CFD constructs the Event and Clip Memories once during offline indexing and reuses them across questions, so the cloud operates on the compact cached index by default and transfers only localized sparse keyframes when textual evidence is insufficient.

Table 15 makes this cost structure explicit, reporting offline captioning time, online serving time, cached-memory size, and GPU energy under the fixed eight-worker setup, using one H200 for Qwen3-VL-8B captioning and one H200 for Qwen3-VL-32B reasoning.

The table separates the one-time offline cost from the per-query online cost. The amortized serving time is $T_{\text{amortized}}(Q) = T_{\text{offline}}/Q + T_{\text{online}}$, where Q is the number of questions sharing the same cached caption memories, T_{offline} is the one-time offline captioning time per video, and T_{online} is the online LLM serving time per query. Including preprocessing, cache writing, frame extraction, and orchestration, the corresponding end-to-end averages are **229.3 s/video offline** and **68.2 s/query**

Table 13: Performance comparison on Video-MME (Fu et al., 2025) across multiple backbone configurations. † denotes reproduced results under strictly identical setups using the authors' code. *#Captions* denotes the average temporal span covered by each caption. *Re-caption?* indicates whether a method revisits video frames at query time to (re)caption or refine its textual memory. *#Frames* counts only raw frames sent to the online answering model.

Method	#Captions	Re-caption?	#Frames	Video-MME			
				short	medium	long	overall
Leading Open-source MLLMs							
VideoChat2-7B (Li et al., 2024b)	-	-	16	48.3	37.0	33.2	39.5
LongVA-7B (Zhang et al., 2024b)	-	-	128	61.1	50.4	46.2	52.6
Kangaroo-7B (Liu et al., 2024)	-	-	64	66.1	55.3	46.6	56.0
Video-CCAM-14B (Fei et al., 2024)	-	-	96	62.2	50.6	46.7	53.2
VideoXL-7B (Shu et al., 2025)	-	-	128	64.0	53.2	49.2	55.5
Dispider-7B (Qian et al., 2025)	-	-	1 fps	-	-	-	57.2
VideoChat-Online-4B (Huang et al., 2025)	-	-	2 fps	-	-	47.1	54.4
TimeChat-Online-7B (Yao et al., 2025)	-	-	1 fps	-	-	48.4	62.4
Qwen3-VL-32B (Bai et al., 2025a)			768	82.4	76.3	69.0	75.9
Agent-based Systems							
VideoAgent (Fan et al., 2024)	2s / caption	✓	-	-	-	46.4	-
VideoTree (Wang et al., 2025d)	8s / caption	✓	-	67.8	59.9	54.2	60.6
DrVideo (Ma et al., 2025)	5s / caption	✓	-	-	-	51.7	-
MemVid (Yuan et al., 2025)	-	✗	1 fps	73.9	63.1	55.0	64.0
VideoLucy (Zuo et al., 2025)†	< 30s / caption	✓	-	73.2	64.7	56.2	64.7
CFD (Qwen3-VL-8B + Qwen3-VL-32B)	233.1s / caption	✗	5.8	72.2	66.6	63.6	67.5
CFD (InternVL3.5-8B + InternVL3.5-38B)	233.1s / caption	✗	7.6	69.0	63.1	60.7	64.3
CFD (Qwen3.5-9B + Qwen3.5-27B)	233.1s / caption	✗	2.5	73.7	70.4	64.1	69.4

Table 14: Online LLM serving time on Video-MME.

Policy	Online LLM serving time per query
No frames	59.6 s
Always frames	63.3 s
Visual-Need Router	68.0 s

online; these end-to-end measurements are distinct from model-serving time. These measurements include server idle and serving overhead and should therefore be interpreted as conservative server-level measurements, not as measurements on a smartphone or wearable device.

Visual-Need Router reliability. Since Video-MME has no routing annotations, we construct an **outcome oracle** from a paired diagnostic rerun over all 2,700 questions. The oracle uses frames only when the always-frame policy corrects an error made by the no-frame policy; it is an empirical upper bound rather than semantic ground-truth annotation. Table 16 summarizes the Router's trigger rates and its coverage of frame-beneficial cases.

Oracle YES means that always-frame is correct while no-frame is wrong; Oracle NO means that no-frame is correct; cases where neither policy is correct are unresolved. Among 1,876 decidable questions, the confusion matrix gives **68.5% routing accuracy** and **95.5% recall** for frame-beneficial questions (Table 17). Table 18 compares the outcome oracle with the no-frame, always-frame, and Router policies across duration splits.

The comparison highlights the remaining failure modes: mixed temporal-perceptual questions, unnecessary routing caused by low textual confidence, and missed evidence after correct routing due to localization or sparse sampling. Confidence calibration, lightweight supervision, and learned routing policies are promising future directions.

Evidence-depth allocation and perceptual diagnostics. The component and budget studies in the main paper establish the overall accuracy–cost trade-off; here we make explicit how the system allocates evidence at different depths. The caption-activation comparison in Table 10 shows that the all-clip baseline supplies all uniformly sampled 30-second Clip Memory captions to every query, whereas CFD activates only captions overlapping localized events. Text-only backtracking cannot recover evidence absent from the offline index, so this selective activation provides local detail without exposing the full clip memory. Event Memory uses TransNetV2 boundaries, while Clip Memory is generated uniformly and independently of event segmentation. The story-first loop then adapts reasoning depth to evidence sufficiency: simple global

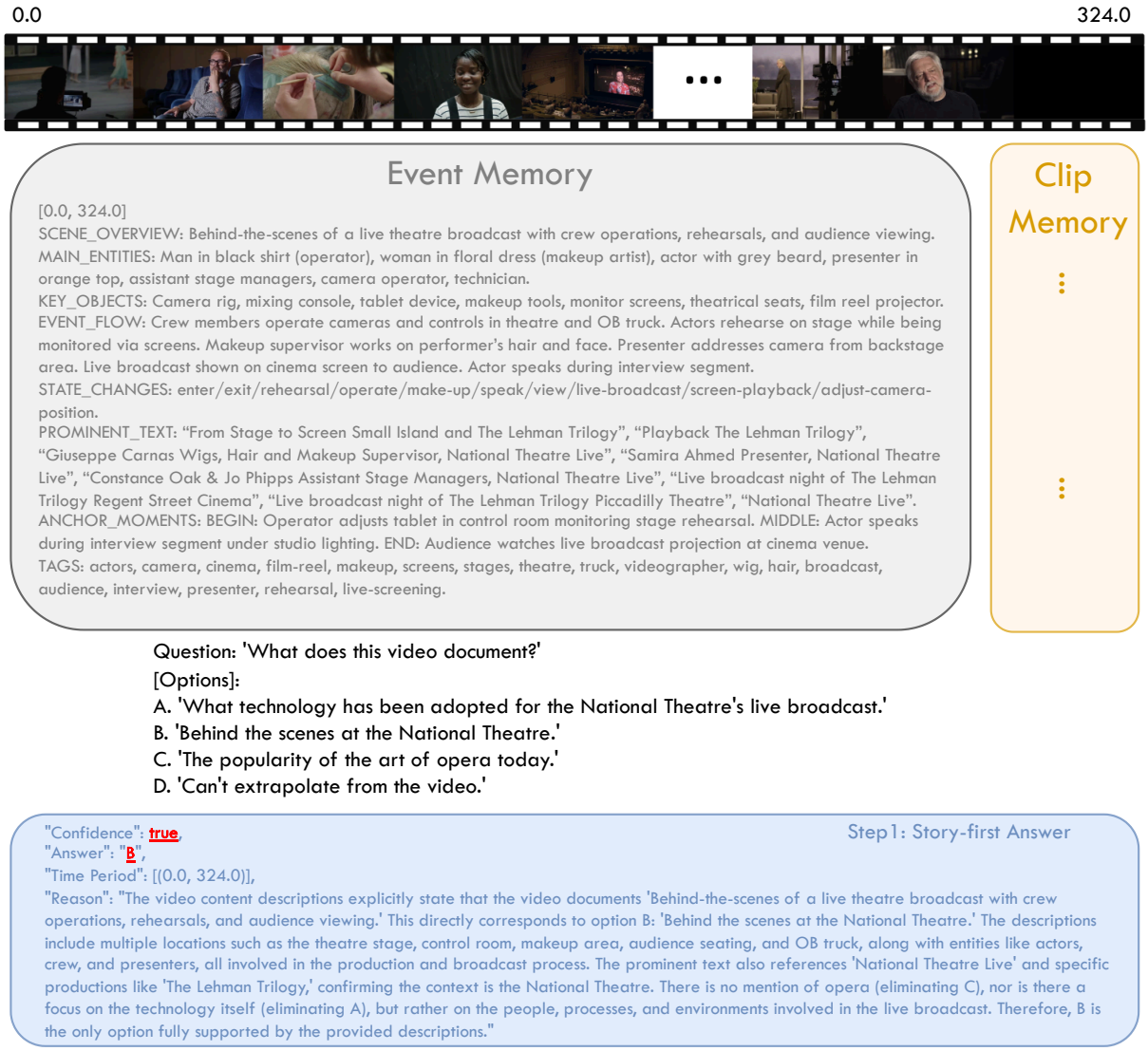

Figure 4: A qualitative visualization of the direct-answer mode of CFD on a Video-MME example. Since the coarse event memory already provides sufficient global semantic evidence, the model answers correctly at the initial story-first stage without temporal backtracking or visual evidence retrieval.

questions can terminate after coarse Event Memory, whereas harder questions progressively invoke temporal localization, localized Clip Memory, and, only when necessary, sparse visual evidence. Figures 4 and 5 illustrate these two evidence-depth regimes.

The complete per-category Router statistics are reported in Table 19; routing is especially frequent for Attribute Perception and OCR, where pixel-level evidence is most likely to be decisive.

Table 20 reports the additional matched diagnostic comparison, in which on-demand pixels improve **76.98**→**82.73** on OCR and **70.37**→**75.93** on Spatial Perception.

F More Implementation Details

Captioning Agent. The offline Captioning Agent runs on the edge device and is invoked once per video in a question-agnostic indexing pass. It operates under two distinct prompt regimes depending on the memory tier being constructed. For *Event Memory*, the agent is prompted to produce a structured event profile covering scene overview, entity listing, chronological event flow, state changes, prominent text, and three retrieval anchor moments. For *Clip Memory*, the agent is prompted to produce plain-sentence micro-action descriptions at high temporal granularity, one observable action per sentence, with explicit instructions on object attribute inclusion, spatial relations, and verbatim

Table 15: Offline/online cost accounting on Video-MME. Offline captioning is one-time per video; online serving is per query. Energy integrates both H200 GPUs over one offline indexing pass and three Video-MME questions per video.

Metric	Short	Medium	Long	Average
Offline captioning (s/video)	34.7	129.1	502.4	222.1
Online LLM serving (s/query)	39.3	66.7	97.9	68.0
Cached caption memory (MB/video)	0.0042	0.0191	0.0826	0.0353
GPU energy (Wh/video + 3 queries)	2.91	7.24	20.53	10.23
Amortized time, Q=1/3/5/10 (s/query)	74.0/50.9/46.2/42.8	195.8/109.7/92.5/79.6	600.3/265.3/198.3/148.1	290.0/142.0/112.4/90.2

Table 16: Router trigger statistics from the paired diagnostic rerun.

Metric	Short	Medium	Long	Average
Queries retrieving frames	34.1%	46.1%	58.3%	46.2%
Frame-beneficial cases captured	42/46 (91.3%)	46/47 (97.9%)	39/40 (97.5%)	127/133 (95.5%)

Table 17: Router decisions against the outcome oracle on the 1,876 decidable questions in the paired diagnostic rerun.

	Router YES	Router NO
Oracle YES	127	6
Oracle NO	585	1,158

Table 18: Outcome-oracle analysis on Video-MME. The oracle is an empirical upper bound, not a deployable policy. Each cell is (accuracy / online frames per query).

Policy	Short	Medium	Long	Average
Outcome oracle	74.0 / 0.33	69.9 / 0.73	64.6 / 0.87	69.5 / 0.64
Visual-Need Router	72.2 / 2.38	66.6 / 5.81	63.6 / 9.09	67.5 / 5.76

Table 19: Category-level Router statistics from the paired diagnostic rerun.

Category	Router YES rate	Frames/query	Accuracy
Attribute Perception	83.57%	4.02	77.93
OCR Problems	86.96%	3.37	82.73
Temporal Reasoning	69.76%	6.45	63.28
Action Reasoning	74.89%	9.51	58.25

Table 20: Matched diagnostic comparison of text-only evidence and on-demand pixels.

Category	Text-only	On-demand pixels
OCR	77.0	82.7
Spatial Perception	70.4	75.9

transcription of visible text. The full prompts are provided in Tables 21 and 22.

Answering Agent. Following prior agent-based video QA systems (Zuo et al., 2025), the Answering Agent receives the current narrative memory assembled from Event Memory with selectively injected Clip Memory entries, and optionally the visual evidence in the FIFO working memory $\mathcal{W}$, assembled into a single multimodal prompt. Given this context and the question, the agent determines whether the available evidence is sufficient to answer with confidence. It outputs a structured dictionary with four fields: `Confidence` (boolean), `Answer` (the selected option letter when confident, `"No Answer"` otherwise), `Time Period` (supporting time intervals when confident, `"No Time"` otherwise), and `Reason` (a chain-of-thought justification that must provide absolute evidential support for the answer). The full prompt is provided in Table 23.

Localization Agent. Following (Zuo et al., 2025), the Localization Agent identifies the single most question-relevant unexplored event from the current narrative memory, excluding all time periods already visited in the current query session. It outputs the selected time period with reasons. The full prompt is provided in Table 24.

Visual-Need Router. After a text-only answer attempt fails to reach confidence, the router receives the selected event, the question, and the Step 3 failure reason, and decides whether sparse keyframes from the current event are likely to improve correctness. The routing decision follows a perceptual-versus-temporal rubric: it returns `YES` when the missing evidence is visual-perceptual in nature, such as object appearance, clothing color, visible text, spatial layout, or fine-grained state cues that cannot be reliably inferred from text descriptions alone; it returns `NO` when the question primarily concerns temporal structure, scene transitions, or long-horizon narrative unfolding, where the textual memory already provides a more suitable abstraction than sparse frames. The full prompt is provided in Table 25.

G Licenses of Datasets and Models

For full transparency and reproducibility, we document the licenses of all datasets, pre-trained models, and software artifacts used in this work. All assets

Table 21: Event Memory captioning prompt (Captioning Agent, offline).

Goal

You are an event-level video indexer. The input is ONE event clip cropped from a longer video. Your output will be used for: (1) retrieval / localization, (2) building a storyline skeleton, (3) deciding what to inspect in finer memory.

Rules

- Describe ONLY what is directly observable in this event clip. Do NOT guess names, roles (*e.g.*, protagonist/antagonist), intent, emotions, relationships, or causes.
- Avoid subjective tone words (*e.g.*, tense, dramatic, emotional, suspicious, implied).
- If something is uncertain, state it as "unclear" rather than guessing.
- Keep it compact: prefer short phrases over long sentences.
- If the event contains multiple distinct scenes/locations, explicitly represent the scene breaks and the chronological order.

Output Format (strict key-value; no extra text)

`SCENE_OVERVIEW:` ⟨1 sentence: dominant setting + time/lighting if visible⟩
`LOCATION_SEQUENCE:`
  - ⟨L1: place/area descriptor⟩
  - ⟨L2: place/area descriptor⟩ ...
`MAIN_ENTITIES:` ⟨comma-separated; include visible attributes: clothing color/type, notable features⟩
`KEY_OBJECTS:` ⟨comma-separated; only salient objects that appear or are interacted with⟩
`EVENT_FLOW` (chronological phases; 4–8 bullets, each is ONE major beat):
  - ⟨Phase 1: who + where + main visible action/change⟩
  - ⟨Phase 2: ...⟩
`STATE_CHANGES:` ⟨opened/closed/on/off/enter/exit/sit/stand/hand-off etc. | none⟩
`PROMINENT_TEXT:` ⟨verbatim visible overlay/sign/subtitle | none | unreadable⟩
`ANCHOR_MOMENTS` (for retrieval; exactly 3):
  - BEGIN: ⟨one short factual snapshot⟩
  - MIDDLE: ⟨one short factual snapshot⟩
  - END: ⟨one short factual snapshot⟩
`TAGS` (max 60; comma-separated): ⟨only concrete nouns/short phrases: people attributes, objects, locations, visible texts, core actions; no full sentences⟩

User turn: Write the event profile now.

are used strictly for non-commercial academic research, in accordance with their respective license terms and intended use.

Datasets. *Video-MME* (Fu et al., 2025) releases its annotations and evaluation code on GitHub under the MIT License, while the underlying videos are sourced from YouTube and are governed by YouTube's Terms of Service; the dataset is explicitly intended for academic, non-commercial research use. *InfiniBench* (Ataallah et al., 2025) releases its annotations under the CC BY-NC-SA 4.0 license, with the underlying movie and TV-show content distributed under the terms of the original sources; usage is limited to non-commercial research. *LVBench* (Wang et al., 2025a) is released under the CC BY 4.0 license; the underlying YouTube videos remain subject to YouTube's Terms of Service. Our use of all three benchmarks is restricted to evaluating long-video question answering in an academic setting, which is consistent with their intended scope.

Multimodal large language models. The *Qwen3-VL* (Bai et al., 2025a), *Qwen2.5-VL* (Bai et al., 2025b), and *Qwen2-VL* (Wang et al., 2024a) series, including the `Qwen3-VL-8B-Instruct` and `Qwen3-VL-32B-Instruct` variants used as our default captioner and reasoner, are released by Alibaba under the Apache License 2.0. The *Qwen3.5* (Qwen Team, 2025) series, used as an alternative backbone in Appendix E, is similarly released under the Apache License 2.0. The *InternVL3.5* (Wang et al., 2025b) family, also used as an alternative backbone, releases its model weights under the Apache License 2.0, with components

Table 22: Clip Memory captioning prompt (Captioning Agent, offline).

1) Task Description You are a precise visual captioner for micro-detail video logging. The input is one short clip cropped from a longer video. Your output will be used for retrieval, evidence alignment, and answering questions later.
2) Instructions - Output ONLY plain sentences. No headers, no bullet points, no structured keys. - Describe ONLY what is directly observable. Do not guess intent, emotions, names, or unseen causes. - POV rule: if the video is clearly first-person POV, use first-person ("I…"); otherwise use third-person ("A person…", "The man…", "The camera…"). - Granularity rule: exactly ONE micro-action per sentence. If nothing moves, write one sentence describing the stable state (who/what/where). - Information density: include object attributes (color/type/shape), spatial relations (left/right/on/in/near/behind), and state changes (opened/closed/on/off) whenever visible. If any text or symbol appears, transcribe it verbatim. If text is present but unreadable, write "text unreadable". - Keep sentences short and factual. Prefer concrete nouns over pronouns when possible.
3) Examples *I am in a kitchen. I pick up a red mug from the table.* *A man stands by a white refrigerator. He opens the refrigerator door.*
User turn: Caption this clip:

inheriting upstream licenses from their base language and vision encoders. *DeepSeek-R1* (Guo et al., 2025), referenced as the agent base model of a comparison method (VideoLucy) on LVBench, is released under the MIT License. Our use of all of the above models is for academic evaluation and does not involve redistribution of the weights.

Software and auxiliary models. *TransNetV2* (Soucek and Lokoc, 2024), used for shot-boundary detection during offline event segmentation, is released under the MIT License. *vLLM* (Kwon, 2025), used to serve both the captioner and the reasoner, is released under the Apache License 2.0. All other supporting libraries used in our pipeline (*e.g.*, PyTorch, Transformers, NumPy) are released under standard permissive open-source licenses (BSD-3-Clause / Apache 2.0 / MIT), and are used in accordance with their respective terms.

We have verified that the licenses and intended uses of the above assets are compatible with academic research use, and that our usage in this paper falls strictly within their permitted scope.

H Use of AI Assistants

Throughout this paper, we used AI assistants for limited assistance with code completion, debugging, and language polishing of the manuscript. All research ideas, experimental design, analysis, and conclusions are the authors' own. No AI assistant was used to generate experimental results or scientific claims.

Table 23: Prompt for the Answering Agent. Given the current narrative memory and optional visual evidence, the agent determines whether sufficient information exists to answer confidently.

The following provides a rough description of what's shown in the video during different time periods: *{Current Memory List (Time Period + Content Description)}* Note that since these descriptions are not very complete and detailed, some key information in the video segments of each time period may not all appear in these content descriptions. Now, a question has been raised regarding the content descriptions of this video. *{Question and Options}* Please read the given video content descriptions and the question in depth, and determine whether you can accurately answer the given question solely based on the currently provided descriptions. If you can answer it with absolute confidence, please answer this question and provide the time periods you are referring to. The answer you provide must have completely and absolutely objective support in the video descriptions. Do not make inferences arbitrarily. Please note that there is only one option that can answer this question. If you think the current content descriptions of the video are still insufficient to accurately answer the question, please do not answer it and give me your reason. You must note that if an ordinal number appears in the provided question, in the vast majority of cases, you should not simply assume that this ordinal number represents the ordinal of the provided time period. You need to focus on understanding the specific meaning of this ordinal number within the question based on all the content descriptions. Output in a strictly standardized dictionary with four key-value pairs: `"Confidence"`: boolean. `True` if certain; `False` if not. `"Answer"`: the answer option letter when confident; `"No Answer"` otherwise. `"Time Period"`: list of `(start, end)` tuples when confident; `"No Time"` otherwise. `"Reason"`: reasoning that absolutely supports the answer. No additional comments should be added within the dictionary.

Table 24: Prompt for the Localization Agent. Given the current narrative memory and the set of already-explored events, the agent identifies the single most question-relevant unexplored time period.

There is currently a video with a total duration of *{video length}* seconds. The following gives a general description of what is shown in the video during certain time periods: *{Current Memory List (Time Period + Content Description)}* Now, a question has been raised regarding this entire video. *{Question and Options}* Please read the given video content descriptions and the question in depth. You do not need to answer this question. Your task is to identify, based on the video content in each time period, the single time period that is most relevant to the question and that you think requires further elaboration of its video content details to make the answer to this question more explicit. Notably, you only need to select the most relevant one from the time periods other than the following already-explored periods: *{Already Searched Time Periods}* Output in a strictly standardized dictionary with two key-value pairs: `"Time Period"`: list containing the single most relevant `(start, end)`. `"Reason"`: justification for the selected time period. No additional comments should be added within the dictionary.

Figure 5: A qualitative visualization of the inference trajectory of CFD on a Video-MME example. The model first reasons over coarse event memory, then progressively narrows the relevant temporal span with clip memory, and invokes sparse visual backtracking only when textual evidence remains insufficient.

Table 25: Prompt for the Visual-Need Router. Given the current selected event, the question, and the Step 3 failure reason, the router decides whether sparse keyframes from the selected event will improve correctness.

You are a routing agent for an agentic multimodal long video QA system. We already attempted answering using text memories (EventMemory + injected ClipMemory captions), but confidence was low. Keyframes are sparse local evidence; captions are better for temporal abstraction (scene transitions, long-horizon unfolding). Now decide whether to extract a few visual keyframes from the **current selected event**. Important: you are not deciding whether video is useful in general; you are deciding whether keyframes from *this* event will likely improve correctness for *this* question.
Current Event: from *{start}*s to *{end}*s. Event Text Memory: *{Event Memory}*
Step 3 Failure Reason: *{Reason}*
Question: *{Question and Options}*

A) Return YES, the missing evidence is visual-perceptual and likely visible in a few keyframes from this event: appearance (clothing, colors, patterns); object attributes (type/shape/color); visible text, signs, overlays, logos; spatial/state cues (left/right, holding, on/in/near, open/closed, on/off); micro-action disambiguation within this event (which hand/tool, exact contact, small gesture). Step 3 reason indicates ambiguity that a look can resolve (cannot distinguish options based on text).
B) Return NO, the question requires temporal abstraction where captions are more suitable: temporal context focus (how things evolve over time, what happens first/then/later, timeline/unfolding, duration-based reasoning); scene transitions (when/where the scene changes, cut/transition/montage, switching locations/activities across a long segment); multi-stage processes within a long span where captions already summarize phases better than a few frames; global narrative linking multiple events or summarization-level questions. Step 3 reason indicates need for broader temporal coverage (earlier/later parts, multiple moments) rather than local visual detail.
C) Keyword heuristics (weak signals): lean YES: color, wearing, looks like, logo, label, text says/reads, number on, left/right, holding, where is the object. Lean NO: scene change/transition/cut, timeline, unfold, sequence, earlier/later, before/after (across moments), how long, over time, montage.
D) Default policy (type-driven): if the question is primarily perceptual-local $\rightarrow$ YES; if primarily temporal-structural / scene-transition $\rightarrow$ NO; if mixed, follow Step 3 failure reason: perceptual uncertainty $\rightarrow$ YES, temporal coverage uncertainty $\rightarrow$ NO.
Output a valid JSON object with exactly four keys:
`"Confidence"`: boolean.
`"Answer"`: `"YES"` or `"NO"`.
`"Time Period"`: `"No Time"`.
`"Reason"`: short justification referencing the rubric (A/B/D).