LONG-HORIZON MOVIE REASONING: A SCALABLE BENCHMARK AND STRUCTURED MEMORY FRAMEWORK

Anonymous authors
Paper under double-blind review

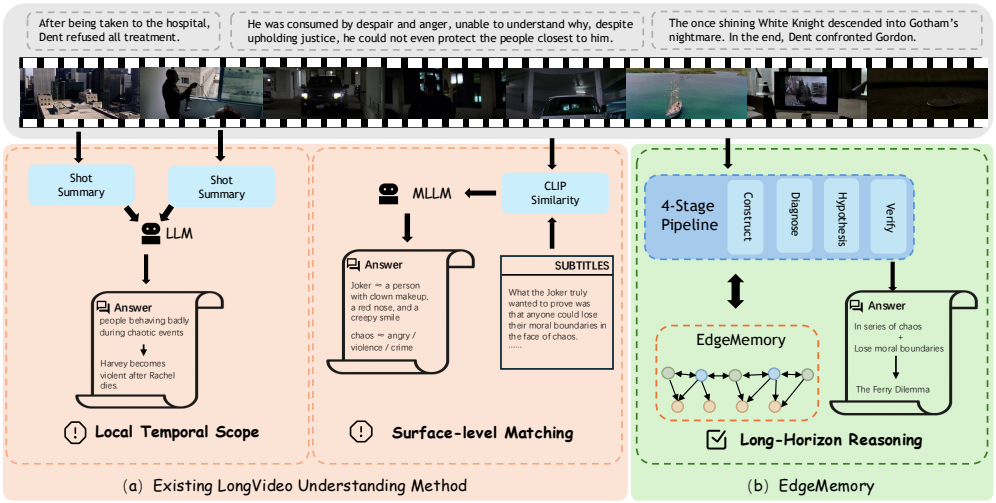


Figure 1: Comparison between the limitations of existing long-video understanding methods and our proposed approach.

ABSTRACT

Long-form video understanding requires models to reason consistently over extended temporal contexts, yet existing benchmarks provide limited evaluation of long-range memory and high-level narrative reasoning. We introduce a semi-automated pipeline for constructing long-form video reasoning benchmarks from full-length movies and their recap narrations, reducing annotation cost while preserving reliable temporal and semantic alignment. Based on this pipeline, we construct **Movie Reason Benchmark (MRBench)**, which evaluates holistic reasoning over long videos with narrative semantics at multiple levels of abstraction. We further introduce complementary metrics for temporal localization under multiple tolerance thresholds and semantic alignment across narrative granularities. To address the challenges exposed by MRBench, we propose **EdgeMemory**, an edge-centric graph-structured memory framework that decomposes complex reasoning tasks, organizes multi-level semantic representations, and progressively refines their relations through a dedicated reasoning pipeline. Extensive experiments show that EdgeMemory consistently improves long-form video reasoning across diverse models. Further analysis reveals distinct capability boundaries and failure modes across model scales, highlighting persistent challenges in long-range evidence integration and abstract narrative reasoning.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have led to significant progress in video understanding (Tang et al., 2025). However, existing tasks primarily focus on relatively

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

short videos (Li et al., 2024; Yang et al., 2025; Törtei et al., 2025) or simple retrieval settings (Wang et al., 2025b; Cao et al., 2025; Fu et al., 2025), lacking systematic evaluation of long-range temporal modeling and complex reasoning capabilities. Meanwhile, movies are a canonical form of long-form video with extended temporal structures and abstract narrative semantics, requiring models to maintain long-term memory and reason over temporally distant events. This raises a critical question: how can we effectively evaluate model capabilities in such realistic scenarios, and more importantly, how can we enhance their ability for long-range multimodal reasoning?

Existing benchmarks fall short of addressing this challenge. Most video benchmarks focus on short clips or question answering tasks that can be solved using only a few key frames (Zhang et al., 2025; Wang et al., 2025a; Liu et al., 2025b), making them inadequate for evaluating long-context reasoning. High performance on needle-in-the-haystack retrieval tasks does not necessarily reflect understanding of video narratives or underlying semantic structures. Meanwhile, existing approaches often rely on relatively simple memory mechanisms (Kim et al., 2025; 2026), such as caching, or retrieve historical content through similarity-based matching with models such as CLIP (Xu et al., 2026; Chen et al., 2025b). While effective for retaining or retrieving locally relevant content, these mechanisms largely treat memory items independently and provide limited support for modeling long-range narrative dependencies across events. As a result, they remain insufficient for complex movie understanding, where grounding often requires reasoning over temporally distant but semantically related evidence.

Motivated by these, we introduce **Movie Reason Benchmark (MRBench)**, a benchmark that aligns high-level recap narratives with their corresponding segments in full-length movies. Unlike conventional video question answering, MRBench uses abstract narrative descriptions rather than localized frame-level queries, requiring models to reason over extended temporal contexts and underlying storyline structure. To enable scalable benchmark construction, we further develop a semi-automatic annotation pipeline that collects movie–recap pairs, extracts and cleans subtitles with Whisper, generates initial temporal alignments through visual feature matching, and refines them via manual verification. Using this pipeline, we collect 141 movies spanning diverse periods, languages, and genres, from which 50 movies are carefully curated and manually verified to form the final benchmark. We evaluate models at multiple narrative granularities, including Beat, Sequence, and Act, with metrics covering both temporal localization and semantic alignment. MRBench presents three key challenges: reasoning over full-length movies, aggregating evidence distributed across distant temporal regions, and bridging the substantial semantic gap between abstract recap narratives and visual content, which requires structured reasoning over characters, events, and their temporal dependencies rather than direct visual–text matching.

Evaluation on MRBench reveals that existing methods struggle particularly with temporally complex and semantically abstract queries, largely due to insufficient long-range evidence integration and shallow cross-modal alignment (Figure 1(a)). These observations suggest that directly reasoning over an entire movie as a monolithic context is insufficient for complex narrative understanding. To address this challenge, we propose **EdgeMemory**, a graph-structured memory framework that first **decomposes the narrative reasoning task into tractable semantic units** and then grounds each unit against structured video memory. Specifically, EdgeMemory represents the task as a set of reasoning plans and organizes video content into append-only semantic nodes, while progressively refining their relationships through a four-stage pipeline: *Construct, Diagnose, Hypothesize, and Verify*. By decoupling stable semantic representations from dynamically updated relations, EdgeMemory supports reusable memory construction, explicit evidence association, and traceable multi-step reasoning, enabling effective long-range narrative understanding without relying on external tools.

In summary, our contributions are as follows:

1. We introduce **Movie Reason Benchmark (MRBench)**, a long-form video benchmark for long-range narrative reasoning, together with a reusable semi-automatic construction pipeline and complementary metrics for temporal localization and narrative alignment.
2. We propose **EdgeMemory**, a structured memory framework that decomposes complex narrative reasoning into tractable semantic units and performs explicit relation modeling through a four-stage *Construct–Diagnose–Hypothesize–Verify* pipeline.
3. We conduct extensive experiments to systematically evaluate the long-context reasoning capabilities of existing models, validate the effectiveness of EdgeMemory, and analyze distinct failure modes across models of different scales.

2 Related Work

2.1 Long Video Understanding

A number of benchmarks have been proposed for long-video understanding, with question answering as the primary evaluation paradigm. MLVU (Zhou et al., 2025) organizes tasks by holistic, single-detail, and multi-detail understanding, while Video-MME (Fu et al., 2025) evaluates both perceptual and reasoning abilities across multiple levels. LVBench (Wang et al., 2025b) further covers six capabilities, including temporal grounding, summarization, reasoning, entity recognition, event understanding, and key information retrieval. LongVideoBench (Wu et al., 2024), in contrast, incorporates interleaved video frames and subtitles and constructs question-answer pairs that require referring and reasoning over long-form content across diverse domains.

Several benchmarks further specialize long-video evaluation to movies. MovieChat (Song et al., 2024) assesses both global comprehension and temporally localized question answering, while MovieBench (Wu et al., 2025) models characters, scene transitions, and narrative progression through a hierarchical structure, with a primary focus on long-video generation. ALLVB (Tan et al., 2025) provides a large-scale QA framework that jointly evaluates understanding, reasoning, and retrieval. Despite their broad coverage, these benchmarks primarily emphasize retrieving or recognizing localized evidence, rather than reasoning over narrative structure.

2.2 Memory-based Frameworks

For long video understanding, recent closed-source models such as GPT-5 (Achiam et al., 2023) and Gemini3 (Team et al., 2023) have achieved strong performance on several tasks, yet they still face two fundamental challenges: dense or uniform frame sampling incurs substantial computational cost, and models often struggle to maintain consistent memory and reasoning states over long temporal horizons. To address these challenges, existing methods mainly explore three complementary strategies: retrieval with vision-language models such as CLIP (Radford et al., 2021), explicit construction of memory banks, and agent-based multi-round retrieval and reasoning.

For example, ELIOT (Liu et al., 2025a) first generates base captions for videos and then extracts structured visual elements, including objects, events, and attributes, for retrieval. OneClip RAG (Chen et al., 2025b) unifies video chunking and cross-modal retrieval-augmented generation (Lewis et al., 2020) within a single processing pipeline, reducing redundant computation. LM2 (Kang et al., 2025) introduces a memory-augmented Transformer architecture, where an external memory module serves as a repository of contextual representations and interacts with input tokens through cross-attention and gated updates. FlexMem (Chen et al., 2026) further treats the KV cache as memory to support effectively longer video understanding. Beyond retrieval and memory design, LVAgent (Chen et al., 2025a) builds a multi-round collaborative agent framework through Selection, Perception, Action, and Reflection, while WorldMM (Yeo et al., 2025) constructs hierarchical memory and uses agents to dynamically search and iteratively refine memory for improved answering.

3 Movie Reason Benchmark

Constructing movie-level benchmarks requires both high-level narrative supervision and precise temporal grounding over hours-long videos, making manual annotation prohibitively expensive. We leverage *movie recap videos*, whose narration provides narrative supervision while synchronized visual clips link it to the original movie timeline. Based on this correspondence, we develop a *recap-driven* semi-automatic pipeline to construct the **Movie Reason Benchmark (MRBench)**.

3.1 Task Definition

We formulate MRBench as a **narrative temporal localization** task for movie-level visual understanding. Given a long movie represented as a sequence of timestamped visual frames, $\mathcal{V} = \{(f_i, t_i)\}_{i=1}^{N}$, and a set of narrative queries extracted from the corresponding movie recap, $\mathcal{S} = \{s_j\}_{j=1}^{M}$, the model is required to localize, for each query s_j, the temporal interval in the original movie that semantically corresponds to the narrated event:

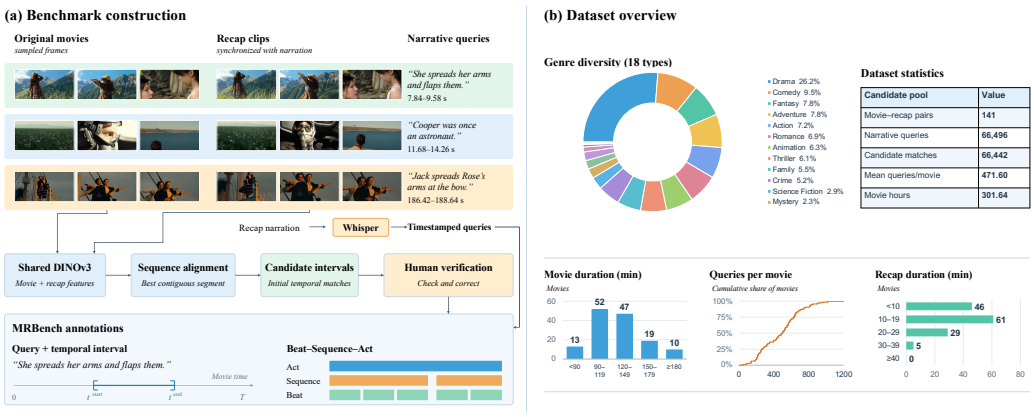


Figure 2: **Overview of MRBench.** (a) Recap clips are aligned with their source movies using shared DINOv3 features and sequence alignment. Whisper timestamps the recap narration, and candidate intervals are manually verified to produce query–interval and beat–sequence–act annotations. (b) The candidate pool contains 141 movie–recap pairs, 66,496 narrative queries, and 66,442 matches across 301.64 hours and 18 genres. The plots summarize the distributions of movie duration, queries per movie, and recap duration.

$$\hat{\mathcal{R}} = \{\hat{r}_j\}_{j=1}^{M}, \qquad \hat{r}_j = [\hat{t}_j^{\text{start}}, \hat{t}_j^{\text{end}}]. \tag{1}$$

Here, f_i denotes the i-th sampled frame in the movie, t_i is its corresponding timestamp, s_j denotes the j-th recap-derived narrative query, and $\hat{r}_j$ is the predicted temporal segment associated.

Compared with conventional moment retrieval, MRBench poses three key challenges: **movie-scale temporal context**, where relevant evidence may span long temporal ranges; **high-level narrative abstraction**, where visual events are expressed at a higher semantic level; and **cross-query dependency**, where queries collectively form a coherent storyline. Together, these properties require models to reason over long-range visual evidence while maintaining narrative consistency across queries.

3.2 Recap-Driven Benchmark Construction

We exploit the natural text–visual correspondence in movie recaps to transform costly manual search and annotation over full-length movies into a semi-automatic process of generation. The pipeline consists of recap-based supervision, cross-video temporal grounding, and human verification with hierarchical narrative annotation, with details provided in the Appendix.

Recap-Based Supervision. Movie recaps provide both high-level narrative descriptions and temporally aligned visual clips. We transcribe the narration using Whisper large-v (Radford et al., 2023) and organize it into timestamped narrative queries. Compared with local dialogue in movie subtitles, recap narration summarizes character actions, event relations, and plot progression from a third-person perspective, providing higher-level narrative supervision. Meanwhile, the visual clip aligned with each query is drawn from the original movie, establishing a natural supervision chain from *narrative query* to *recap visual clip* to *original movie*, which serves as a visual bridge for temporal grounding.

Cross-Video Grounding. We leverage this visual bridge to automatically ground each narrative query in the original movie. Specifically, we encode the recap clips and sampled movie frames with a pretrained DINOv3 encoder and perform sequence-level visual matching to identify the best-matching contiguous segment along the movie timeline. The resulting segment serves as the candidate temporal annotation. By exploiting direct visual correspondence, this formulation avoids the substantially harder problem of matching abstract narrative text directly against hours-long movies.

Human Verification and Narrative Annotation. Because recap editing and visually similar scenes can introduce alignment errors, we manually verify and correct the automatically generated candidates and use the fine-grained verified subset as the official evaluation set. Beyond temporal annotations, we

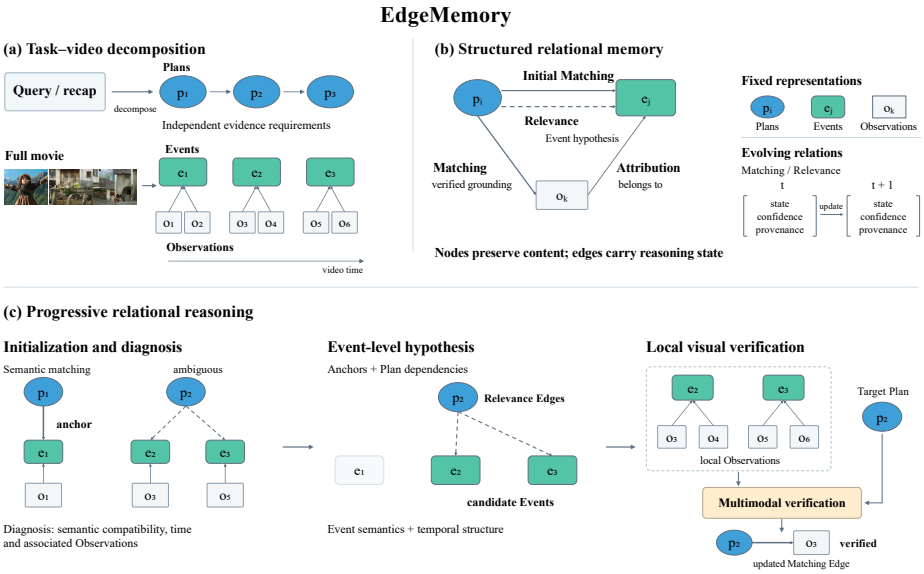


Figure 3: **EdgeMemory.** Overview of our hierarchical long-video reasoning framework.

further organize each movie into a three-level **Beat–Sequence–Act** narrative hierarchy, associating each query with both its temporal interval and narrative units at multiple granularities. This enables evaluation of both temporal localization and narrative consistency.

3.3 Benchmark Statistics

Our construction pipeline yields **141 movie–recap pairs** as candidate data, from which **50 movies** are carefully human-verified to form the official MRBench evaluation set. The verified movies range from **44 to 260 minutes**, with an average of **471.60 narrative queries** per movie, spanning diverse languages, genres, and release years. Fig. 2 summarizes additional dataset statistics and distributions.

3.4 Evaluation Metrics

We evaluate MRBench **temporal localization** and **hierarchical narrative alignment**.

Temporal Localization. For each query, we measure whether the predicted timestamp $\hat{t}_i$ falls within a tolerance δ of the center of its ground-truth interval $[t_i^{\text{start}}, t_i^{\text{end}}]$:

$$\text{Acc@}\delta = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\left|\hat{t}_i - \frac{t_i^{\text{start}} + t_i^{\text{end}}}{2}\right| \le \delta\right), \tag{2}$$

where $\delta \in \{10, 30, 60, 120, 180, 300\}$ seconds denotes the temporal tolerance.

Hierarchical Narrative Alignment. To complement timestamp-based localization, we evaluate whether the prediction falls within the correct narrative unit: **Beat–Sequence–Act** hierarchy:

$$\text{Acc}^{(l)} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(u_i^{\text{pred},l} = u_i^{\text{gt},l}\right), \qquad l \in \{\text{Beat}, \text{Sequence}, \text{Act}\}, \tag{3}$$

where $u_i^{\text{pred},l}$ and $u_i^{\text{gt},l}$ denote the predicted and ground-truth narrative units at level l, respectively.

4 Method

Complex narrative understanding is not merely a long-video retrieval problem. A full-length movie contains multiple narrative events spanning long temporal ranges, while the query or recap may itself

consist of multiple units that must be grounded individually while preserving their dependencies. Thus, effective reasoning requires jointly structuring both the task and the video.

We propose EDGEMEMORY, a structured memory framework that decomposes the task into **Plans** and the video into high-level **Events** with fine-grained **Observations**. Rather than repeatedly reasoning over the full video, EDGEMEMORY progressively establishes and updates relations among these units. Most long-range reasoning is performed over compact semantic representations.

4.1 EDGEMEMORY OVERVIEW

Given an input task, we decompose it into a set of minimally groundable Plans, $\mathcal{P} = \{p_i\}_{i=1}^{N_P}$, where each Plan specifies an independent evidence requirement. The video is represented hierarchically by Events and Observations, $\mathcal{E} = \{e_j\}_{j=1}^{N_E}$, $\mathcal{O} = \{o_k\}_{k=1}^{N_O}$, where Events summarize coherent narrative units and Observations retain visual evidences.

EDGEMEMORY connects these structures through explicit relations. Initial visual matching produces candidate Plan–Observation correspondences. Reliable matches are retained as anchors, while unresolved Plans are reasoned over at the Event level using existing anchors, dependencies among Plans, and Event semantics and temporal structure. The model then verifies only the Observations within selected Events. This shifts most long-range reasoning to a compact semantic space and restricts expensive multimodal inference to a small set of candidate regions.

4.2 STRUCTURED RELATIONAL MEMORY

EDGEMEMORY represents its memory using three types of nodes, $\mathcal{M} = \mathcal{P} \cup \mathcal{E} \cup \mathcal{O}$, corresponding to Plans, Events, and Observations. Plans preserve task semantics and their contextual order, Events provide narrative representations, and Observations retain fine-grained visual evidence for grounding.

The reasoning state is represented by three types of relations:

$$o_k \xrightarrow{\text{attr}} e_j, \qquad p_i \xrightarrow{\text{match}} o_k, \qquad p_i \xrightarrow{\text{rel}} e_j. \tag{4}$$

An **Attribution Edge** organizes Observations under Events; a **Matching Edge** represents direct Plan–Observation grounding; and a **Relevance Edge** represents a Plan–Event hypothesis that constrains subsequent verification. While node representations remain fixed, these reasoning relations are progressively created, evaluated, and updated during inference. Each dynamic relation additionally maintains its state, confidence, and provenance, with detailed schemas provided in the Appendix.

This formulation separates semantic content from reasoning state: nodes preserve stable representations, while inference proceeds through **edge evolution**.

4.3 PROGRESSIVE RELATIONAL REASONING

EDGEMEMORY resolves Plan grounding through three main steps: initialization and diagnosis, event-level hypothesis, and local visual verification.

Initialization and Diagnosis. During initialization, video decomposition establishes **Observation–Event Attribution Edges**, while semantic matching establishes preliminary **Plan–Event Matching Edges** between each Plan and potentially relevant Events. Rather than forcing a single grounding, EDGEMEMORY retains multiple plausible Plan–Event correspondences when the evidence is ambiguous. The model then diagnoses these candidate matches using semantic compatibility, temporal information, and the fine-grained Observations associated with each Event. Reliable matches are retained as **anchors**, while ambiguous or conflicting Plans remain unresolved. Subsequent reasoning therefore focuses only on uncertain Plans, with established anchors serving as contextual constraints.

Event-Level Hypothesis. For an unresolved Plan p_i, EDGEMEMORY identifies a small set of candidate Events, $p_i \to \{e_{j_1}, \ldots, e_{j_K}\}$, conditioned on existing anchors, dependencies among Plans, and the semantic and temporal structure of Events. These Plan–Event Relevance Edges restrict subsequent grounding to plausible narrative regions. Since this stage operates over textual Plan and Event representations, long-range narrative reasoning can be performed in the textual semantic space without repeatedly invoking expensive multimodal inference over the raw video.

Local Visual Verification. Given the candidate Events, EDGEMEMORY accesses only their associated Observations and performs localized multimodal verification with the target Plan. The result updates the corresponding Plan–Observation Matching Edges and determines the final grounding. This design confines expensive visual reasoning to a small number of candidate regions: Events support global narrative reasoning, while Observations provide the fine-grained evidence.

5 EXPERIMENTS

5.1 EXPERIMENTS SETUP

Dataset. We conduct experiments on a 50-movie subset of the proposed benchmark, covering diverse genres, narrative structures, and durations to provide a comprehensive testbed for long-form video understanding and subtitle-to-movie alignment.

Baselines. We evaluate both proprietary and open-source MLLMs, including GPT-5.4 (Achiam et al., 2023), Gemini 3 Pro (Team et al., 2023), and Qwen3-VL (Bai et al., 2025), and compare our method against representative long-video understanding paradigms, including naive sliding-window VLMs, the agent-based framework WorldMM (Yeo et al., 2025), and retrieval-based One-CLIP RAG (Chen et al., 2025b), collectively covering naive matching, memory reasoning, agent-based modeling, and retrieval-based approaches.

5.2 MAIN RESULTS UNDER TEMPORAL THRESHOLDS

Table 1: Comparison of temporal localization accuracy under different thresholds and event-level narrative alignment.

Method	Temporal Accuracy (s)						Event-level Alignment		
	10	30	60	120	180	300	Beat	Sequence	Act
Comparison with state-of-the-art methods									
GPT-5.4	16.43	27.10	33.31	37.90	39.84	41.97	14.69	36.76	45.46
Gemini3 Pro	15.09	24.91	30.00	32.26	34.15	36.60	10.10	30.57	42.45
Qwen3-VL-8B (Bai et al., 2025)*	3.27	6.11	8.77	11.59	13.99	17.50	3.27	13.41	31.41
WorldMM(Qwen3-VL-8B*) (Yeo et al., 2025)	5.92	8.82	10.38	13.51	14.85	19.64	5.80	14.51	33.04
WorldMM(GPT-5.4) (Yeo et al., 2025)	18.00	29.27	35.64	40.18	42.18	43.09	14.18	38.18	46.55
One-CLIP RAG(Qwen3-VL-8B*) (Chen et al., 2025b)	2.65	5.62	8.26	11.19	13.13	16.40	2.13	11.98	27.31
Ours with Different Model									
EdgeMemory (Qwen3-VL-8B*)	4.27	9.45	11.02	16.93	18.11	24.02	6.69	23.23	38.66
EdgeMemory (GPT-5.4)	**19.07**	**33.79**	**43.91**	**52.52**	**56.23**	**60.48**	**18.19**	**50.78**	**67.51**
EdgeMemory (Gemini3 Pro)	17.76	30.61	37.62	39.95	40.42	53.14	14.72	36.21	41.12

* Due to context length limitations, this model require additional strategies such as input chunking.

Table 1 presents a quantitative comparison between EdgeMemory and all baselines. We observe that even state-of-the-art proprietary models struggle to achieve strong performance on this task, highlighting its inherent difficulty. EdgeMemory consistently achieves the best results across both proprietary and open-source model settings, demonstrating its effectiveness and generality.

We further observe a clear performance gap between small-scale and large-scale models, which can be attributed to differences in model capacity and architectural design. Smaller models consistently underperform, and improvements brought by different frameworks remain limited in this regime. As discussed in the section 5.5, the root cause lies in the insufficient multimodal reasoning ability and limited long-context capacity of smaller models.

5.3 EVENT-LEVEL NARRATIVE ALIGNMENT

From Table 1, we observe that larger models consistently outperform smaller ones across all hierarchy levels (Beat, Sequence, and Act). In contrast, the improvements brought by our method over baselines are more pronounced at the Sequence and Act levels. This observation aligns well with our design intuition. The advantages of large-scale models primarily stem from stronger multimodal representations and extended long-context reasoning capacity, leading to uniform gains across all levels. In comparison, our framework preserves the underlying model's fine-grained capability at

Table 2: Ablation study on the four-stage reasoning pipeline of EdgeMemory.

Method	Construct	Diagnose	Hypothesis	Verify	Acc@30↑	Err@30↓	MatchRate↑
Construct only	✓	×	×	×	26.29	58.53	63.38
+ Diagnose	✓	✓	×	×	25.30	52.57	53.34
+ Hypothesis	✓	✓	✓	×	36.02	63.98	–
+ Verify(Full)	✓	✓	✓	✓	38.19	54.76	84.42
w/o Diagnose	✓	×	✓	✓	34.03	62.50	88.37
w/o Hypothesis	✓	✓	×	✓	27.68	63.83	76.52

the Beat level while further enhancing higher-level semantic reasoning and long-range structured memory, resulting in more significant improvements at the Sequence and Act levels. Moreover, our method consistently improves performance even when applied to state-of-the-art proprietary large models, demonstrating its strong scalability and generality.

5.4 Ablation Study

We conduct systematic ablation studies to analyze both the structural components of our method and the contributions of each reasoning stage in a subset of MRBench. **Nodes and Edges Ablation.** We perform three node-level ablation experiments. First, in the Plan ablation, we remove the high-level planning nodes and directly perform reasoning on raw subtitles. Second, in the Event ablation, all Plans directly interact with low-level frames without the intermediate event abstraction layer. Third, in the Observation ablation, reasoning is performed solely over Plan and Event nodes without accessing low-level observational evidence. As shown in Table 3, removing any node type leads to a significant performance drop, demonstrating that the three node hierarchies provide complementary and indispensable functionalities for long-video reasoning.

Stages Ablation. We further analyze the effects of different reasoning stages on accuracy, error rate, and coverage throughout the pipeline. Notably, when the Verify stage is removed, we use the temporal midpoint of the associated Event as the final predicted timestamp. We then ablate the two intermediate stages individually. The results indicate that the Diagnose stage primarily affects the error rate by filtering anomalous matches and reducing noisy associations, whereas the Hypothesis stage mainly influences the overall accuracy by establishing more reliable high-level semantic associations. Detailed results are presented in Table. 2.

Table 3: Ablation on EdgeMemory components.

Method	P	E	O	P-O	E-O	P-E	A@30
Full	✓	✓	✓	✓	✓	✓	38.19
w/o Plan	×	✓	✓	–	✓	–	28.87
w/o Event	✓	×	✓	✓	–	–	26.87
w/o Obs.	✓	✓	×	–	–	✓	30.74

5.5 Qualitative Analysis

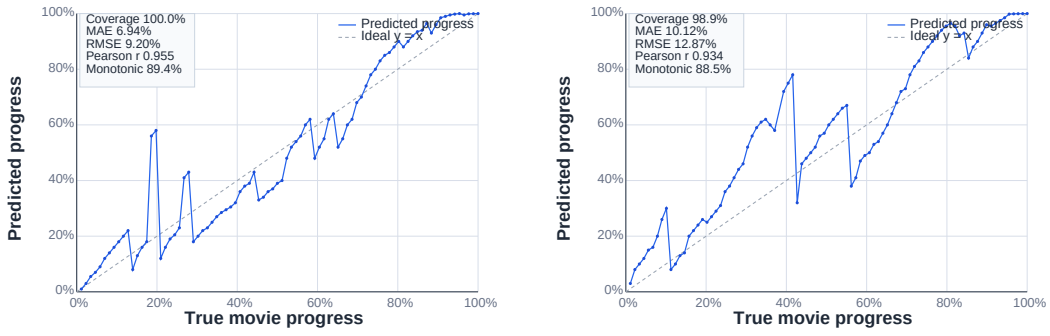
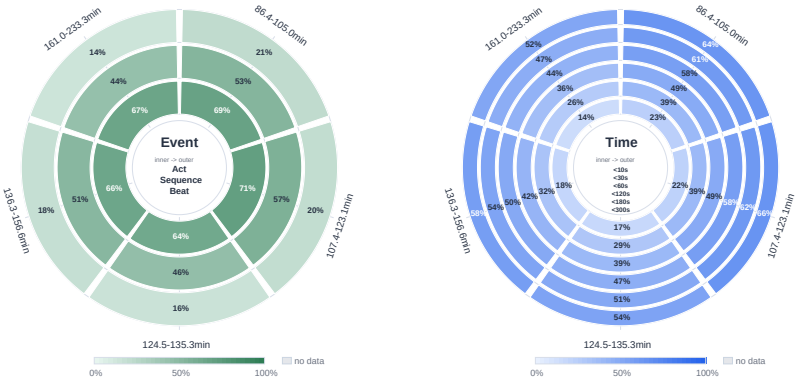


Figure 4: Temporal cognition tracking curves.The model demonstrates an overall awareness of its temporal position, yet remains susceptible to interference from local semantic cues.



(a) Event Accuracy by Duration (b) Time Accuracy by Duration

Figure 5: Event and Time Accuracy

Temporal Cognition. We prompt MLLMs to infer the temporal position of input video segments within the overall movie timeline, aiming to analyze their temporal cognition ability under long-form video understanding. Figure 4 presents two representative cases. We observe that models generally maintain a stable and globally consistent sense of narrative progression throughout the movie. Although occasional temporal deviations occur, the models are often able to gradually correct these offsets in subsequent reasoning steps. Overall, MLLMs exhibit a reasonably robust awareness of relative temporal positioning in long-form narratives.

Error Taxonomy. Errors made by MLLMs in long-video understanding exhibit clear hierarchical structures rather than random distributions. Through manual analysis, we categorize failures into three reasoning levels: Perception, Understanding, and Reasoning. Perception errors arise from incorrect visual recognition and are relatively rare due to the strong local perceptual capability of current MLLMs. Understanding errors involve incorrect interpretation of visual or textual semantics, where models still show noticeable failure rates despite strong overall performance. Reasoning errors, including abstract narrative and contextual failures, require integrating distributed evidence and reconstructing implicit narrative logic, making them substantially more challenging for current models. In addition, sparse video sampling introduces Sample Errors, where critical evidence is never observed, leading to unavoidable failures.(Details in A.3.)

Impact of Long-Context Complexity. We further analyze the impact of context length on model performance. As shown in Figure 5a, performance on Act- and Sequence-level alignment gradually declines as movie duration increases, while Beat-level performance remains relatively stable. This suggests that as temporal spans grow longer, high-level narrative structure modeling and reasoning become the primary bottlenecks, gradually approaching the limits of current MLLMs. Similarly, with increasing movie length, accuracy under coarse-grained temporal thresholds degrades progressively, whereas fine-grained temporal localization remains comparatively stable, as illustrated in Figure 5b. These observations indicate that models are still capable of maintaining local temporal grounding, but struggle to preserve globally coherent narrative understanding over extended contexts.

6 Conclusion

In this work, we introduce **MRBench**, a benchmark for long-range narrative reasoning over full-length movies, together with a **semi-automatic pipeline** for constructing complex narrative grounding annotations from movie recaps. To address the challenges revealed by MRBench, we propose **EdgeMemory**, which structures task requirements and video content through Plans, Events, and Observations for explicit relational reasoning. Experiments across multiple model scales show consistent improvements, particularly on higher-level Sequence and Act reasoning. Our results further highlight that coherent long-range narrative reasoning remains a key challenge for current MLLMs, motivating more structured memory mechanisms for long-form video understanding.

AI USE STATEMENT

In this work, we used generative AI tools to assist with data annotation during benchmark construction, to improve the clarity and readability of the manuscript, and to assist with modifications and debugging of existing experimental code. For data annotation, generative AI was used to produce or structure preliminary annotations, which were subsequently reviewed and, where necessary, corrected by human annotators before being included in the final benchmark. Generative AI tools were not used to formulate or refine research hypotheses, design the proposed methodology or experiments, analyze experimental results, or draw scientific conclusions. All AI-assisted text was carefully reviewed and revised by the authors, and all AI-assisted code modifications were manually inspected and tested for correctness before use in the reported experiments. We take full responsibility for the final content of this work, including the benchmark annotations, text, claims, code, and experimental results.

ETHICS STATEMENT

MRBench is constructed from publicly accessible movies and movie-recap videos. We recognize that public accessibility does not imply permission to redistribute the underlying copyrighted audiovisual content. Accordingly, we do not redistribute original movies, recap videos, video clips, extracted frames, or other copyrighted audiovisual material. The planned benchmark release is limited to distributable research artifacts, including annotations, metadata, evaluation and data-construction code, and source identifiers or acquisition instructions where permitted by the corresponding platform policies. Users are responsible for obtaining the underlying source videos through the original platforms or other authorized channels. The final release scope will follow applicable platform terms, copyright requirements, and institutional legal and compliance review. Human annotation in the benchmark construction process is limited to verifying and refining benchmark annotations and does not involve the collection of personal or sensitive information.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce the proposed benchmark and method throughout the main paper and appendix. The benchmark construction procedure, evaluation protocol, and metrics are described in the benchmark and experimental sections. The design and reasoning procedure of EdgeMemory are specified in the method section, while additional implementation details, prompts, hyperparameters, benchmark validation procedures, and efficiency settings are provided in the appendix. We also document the benchmark release scope and data-acquisition procedure to support reconstruction of the evaluation setting without redistributing copyrighted audiovisual content. Evaluation and data-construction code, together with the releasable benchmark annotations and metadata, will be made available as part of the benchmark release.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Zongsheng Cao, Yangfan He, Anran Liu, Jun Xie, Feng Chen, and Zhepeng Wang. Tv-rag: A temporal-aware and semantic entropy-weighted framework for long video retrieval and understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9071–9079, 2025.

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

Tao Chen, Shaobo Ju, Qiong Wu, Chenxin Fang, Kun Zhang, Jun Peng, Hui Li, Yiyi Zhou, and Rongrong Ji. Towards effective and efficient long video understanding of multimodal large language models via one-shot clip retrieval. *arXiv preprint arXiv:2512.08410*, 2025b.

Tao Chen, Kun Zhang, Qiong Wu, Xiao Chen, Chao Chang, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Scaling the long video understanding of multimodal large language models via visual memory mechanism. *arXiv preprint arXiv:2603.29252*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Jikun Kang, Wenqi Wu, Filippos Christianos, Alex James Chan, Fraser David Greenlee, George Thomas, Marvin Purtorab, and Andrew Toulis. Lm2: Large memory models for long context reasoning. In *Workshop on Reasoning and Planning for Large Language Models*, 2025.

Donghyuk Kim, Sejeong Yang, Wonjin Shin, and Joo-Young Kim. V-rex: Real-time streaming video llm acceleration via dynamic kv cache retrieval. In *2026 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 1–14. IEEE, 2026.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. *arXiv preprint arXiv:2506.15745*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Xuye Liu, Yimu Wang, and Jian Zhao. Eliot: Zero-shot video-text retrieval through relevance-boosted captioning and structural information extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 381–391, 2025a.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv e-prints*, pp. arXiv–2503, 2025b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. Allvb: All-in-one long video understanding benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7211–7219, 2025.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Brigitta Malagurski Törtei, Yasser Dahou, Ngoc Dung Huynh, Wamiq Reyaz Para, Phúc H Lê Khac, Ankit Singh, Sofian Chaybouti, and Sanath Narayan. Visres bench: On evaluating the visual reasoning capabilities of vlms. *arXiv preprint arXiv:2512.21194*, 2025.

Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Liu, Jose M Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal grounding. *arXiv preprint arXiv:2507.13353*, 2025a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025b.

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

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28984–28994, 2025.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Long video understanding with learnable retrieval in video-language models. *IEEE Transactions on Multimedia*, 2026.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. *arXiv preprint arXiv:2512.02425*, 2025.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025.

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

A Appendix

A.1 Data Release, Licensing, and Copyright Compliance

MRBench is constructed from publicly accessible movies and movie-recap videos. However, public accessibility does not imply that the underlying audiovisual content is licensed for redistribution. To avoid redistributing copyrighted material, **MRBench will not include or redistribute any original movie files, recap videos, movie clips, extracted video frames, or other copyrighted audiovisual content from the source videos**.

Instead, the planned public release will contain only benchmark artifacts that can be distributed independently of the underlying copyrighted videos, including:

- evaluation and data-construction code;
- derived temporal annotations and benchmark metadata;
- source identifiers and corresponding data-acquisition instructions, where such information can be provided under the applicable platform terms and release policy.

Users will therefore be responsible for obtaining the underlying source videos through the original platforms or other authorized channels. The released benchmark metadata and annotations are intended to enable reconstruction of the evaluation setting without distributing copies of the original copyrighted audiovisual content.

To make the provenance and release conditions of the benchmark transparent, we will additionally provide a **data asset sheet** for the released components. For each data source or benchmark component, the asset sheet will document, where available, its source, ownership or rights-holder information, relevant platform terms, authorization status, and intended form of release. This documentation is intended to clearly distinguish publicly accessible source material from artifacts that the benchmark itself is permitted to redistribute.

The final release scope and distribution procedure will be determined only after institutional legal and compliance review. If a source or derived component is determined not to be suitable for redistribution, it will not be included in the public release. Accordingly, the benchmark release should be understood as a release of *annotations, metadata, and evaluation infrastructure*, rather than a redistribution of the underlying movies or recap videos. “‘

A.2 Implementation Details.

All open-source models are deployed locally and executed on four NVIDIA RTX 4090 GPUs, while proprietary models (GPT-5.4 and Gemini 3 Pro) are accessed via their official APIs. To ensure fairness, all methods are evaluated under the same input constraints, including a fixed upper bound on the number of visual inputs per inference step (e.g., up to 50 frames per call), and follow identical evaluation protocols.

The four-stage pipeline formulates reasoning as iterative edge evolution: Construct initializes coarse relations, Diagnose identifies unreliable edges, Hypothesis introduces event-level constraints, and Verify refines local alignments within constrained temporal windows. Such iterative refinement is essential for handling noisy grounding, long-range dependencies, and limited perception in MLLMs.

A.3 More details about quality analysis

Errors made by MLLMs in long-video understanding are not randomly distributed, but instead exhibit clear hierarchical structures. As illustrated in Figure 6, we categorize model failures into three reasoning levels and conduct manual error analysis on a subset of the evaluation data. The first level corresponds to Perception, which mainly consists of perception errors where the model fails to correctly recognize visual content. Current large-scale MLLMs already demonstrate strong capability in local multimodal perception tasks, with only a small number of errors remaining in smaller models. Overall, performance at this level is close to human-level accuracy, resulting in relatively low error rates. The second level corresponds to Understanding, including Visual Interpretation Errors and Text Interpretation Errors. This level requires the model to correctly interpret visual content and textual semantics, representing the core competency required by many existing multimodal tasks and the primary capability range of current MLLMs. Although models can successfully solve a substantial portion of such cases, noticeable failure rates still remain. The third level corresponds to Reasoning, including Text Interpretation Errors, Abstract Narrative Errors, and Contextual Errors. This level requires joint reasoning over multiple distributed pieces of evidence and the ability to reconstruct concrete scenes from abstract narrative descriptions. Compared with the previous two levels, this setting is substantially more challenging, and current models still struggle to solve such cases reliably and consistently. In addition, sparse video sampling introduces an additional category of Sample Errors, where critical evidence is never observed by the model, leading to unavoidable reasoning failures.

A detailed analysis of all error types is shown in Figure 7.

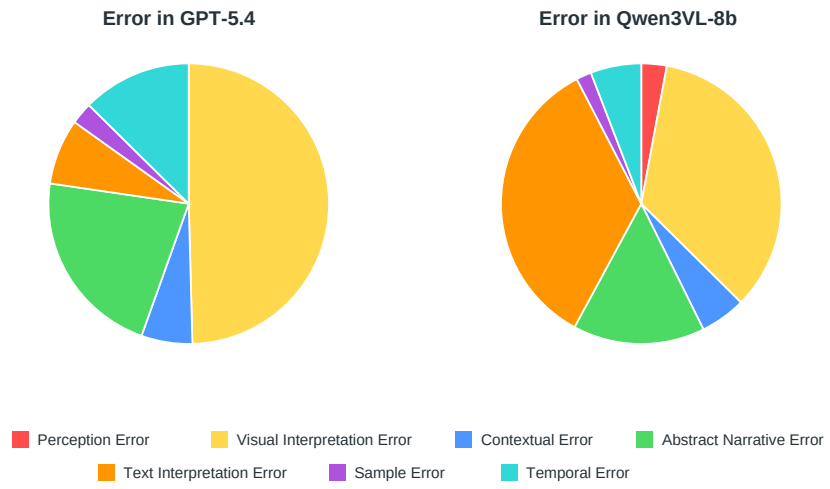


Figure 6: Error in Different Model. Under long-context settings, smaller models exhibit occasional perceptual errors and weaker cross-modal reasoning ability, while larger models remain more robust.

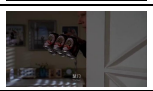
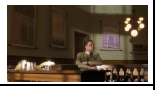
Error Type		Subtitle	Prediction	Ground-Truth	Reason & Analysis
Sample Error	Sample Error	He disguises himself as a doctor.			Reason (GPT): The silhouette of a doctor directly alludes to the act of impersonating a doctor. Analysis (Human): no sampled ground-truth frame accurately matches the subtitle text.
Perception Error	Perception Error	At this moment, Marlowe comes to Truman with beer in his hand.			Reason (GPT): The frame shows Marlowe appearing at Truman's house holding a beer. Analysis (Human): The model fails to perceive key visual content (beer) in the frame.
Interpretation Error	Text Interpretation Error	Disguise himself as the newly arrived substitute teacher.			Reason (GPT): Frank stands at the courtroom with a confident posture, as if assuming an authoritative role. Analysis (Human): The model misinterprets the meaning of the subtitle text.
	Visual Interpretation Error	Frank's letter to his father.			Reason (GPT): The desktop area inside the room best matches the "discovering the letter". Analysis (Human): The model misunderstands the semantic meaning of the visual frame.
Reasoning Error	Contextual Error	Christmas Eve, 1969.			Reason (GPT): Carl working late in his office, with Christmas tree decorations placed beside his desk. Analysis (Human): The model fails to use surrounding subtitles "Carl meets Frank." for matching.
	Temporal Reasoning Error	It must have slipped down from your neck.			Reason (GPT): The caption of this frame corresponds directly to the line. Analysis (Human): The model selects a similar event from the wrong time segment.
	Abstract Narrative Error	I'll have everything all to myself.			Reason (GPT): This close shot emphasizes the father's emotional pause, and also follows up on Frank's hopeful expression about his future life. Analysis (Human): The model fails to infer the concrete event behind abstract narrative text.

Figure 7: All Error Types

Errors made by MLLMs in long-video understanding are not randomly distributed, but instead exhibit clear hierarchical structures. As illustrated in Figure 6, we categorize model failures into three reasoning levels and conduct manual error analysis on a subset of the evaluation data. The first level corresponds to Perception, which mainly consists of perception errors where the model fails to correctly recognize visual content. Current large-scale MLLMs already demonstrate strong capability in local multimodal perception tasks, with only a small number of errors remaining in smaller models. Overall, performance at this level is close to human-level accuracy, resulting in relatively low error rates. The second level corresponds to Understanding, including Visual Interpretation Errors and Text Interpretation Errors. This level requires the model to correctly interpret visual content and textual semantics, representing the core competency required by many existing multimodal tasks and the

primary capability range of current MLLMs. Although models can successfully solve a substantial portion of such cases, noticeable failure rates still remain. The third level corresponds to Reasoning, including Text Interpretation Errors, Abstract Narrative Errors, and Contextual Errors. This level requires joint reasoning over multiple distributed pieces of evidence and the ability to reconstruct concrete scenes from abstract narrative descriptions. Compared with the previous two levels, this setting is substantially more challenging, and current models still struggle to solve such cases reliably and consistently. In addition, sparse video sampling introduces an additional category of Sample Errors, where critical evidence is never observed by the model, leading to unavoidable reasoning failures.



Figure 8: Qualitative and Failure Analysis. Different failure patterns between large-scale and small-scale models

Case Study. During evaluation, we observe fundamentally different failure patterns between large-scale and small-scale models. For large models, the most common failures arise from matching semantically similar yet narratively different events, or from excessive reasoning behaviors such as concept over-binding. In contrast, under long-context pressure, smaller models tend to degenerate into repetitive reuse of previously observed evidence or simplistic subtitle-driven matching patterns, as illustrated in Figure 8. At the reasoning level, even when predictions are incorrect, large models still exhibit attempts at narrative-level reasoning and semantic association, whereas smaller models largely remain at the stage of local visual similarity matching. These observations suggest distinct research directions for different model scales. For large models, future efforts should focus on developing more effective reasoning frameworks and constraint mechanisms to better regulate and utilize their high-level reasoning capabilities. For smaller models, improvements are more likely to come from stronger context compression strategies and auxiliary mechanisms that enhance their fundamental long-context understanding ability.

A.4 Four-Stage Implementation Details

We provide implementation details of the four-stage reasoning pipeline introduced in Sec. 4. Each stage operates on the same structured memory graph but performs a different type of relation update. Construct initializes the graph, Diagnose identifies unreliable grounding relations, Hypothesis introduces Plan–Event structural constraints, and Verify revisits local visual evidence to finalize task-specific outputs.

A.4.1 CONSTRUCT STAGE

The Construct stage initializes the structured memory graph from the task input and the video. Its outputs include Event nodes, Observation–Event Attribution Edges, and initial Plan–Observation Matching Edges.

Streaming observation construction. Given a temporally ordered video, we process frames in consecutive chunks and instantiate the corresponding Observation nodes. Each Observation retains the visual evidence and its temporal position, while previously constructed nodes remain in memory as the video stream progresses.

Joint event construction and initial grounding. For each incoming video chunk, the model jointly performs two operations in a single multimodal inference step. First, it abstracts the current visual content into one or more Event nodes that summarize the corresponding narrative content. Second, it matches active Plans against the Observations in the current chunk and establishes candidate Plan–Observation Matching Edges.

To support long recap sequences without presenting all Plans at every step, we maintain a dynamic pool of active Plans. A Plan remains active while its corresponding narration is still being grounded. Once sufficient evidence has been associated with the Plan, it is removed from the active pool and replaced by subsequent Plans. This allows the model to process the video sequentially while maintaining only the currently relevant task units.

For each constructed Event, we additionally establish Observation–Event Attribution Edges that record which Observations support the Event. The Construct stage therefore produces an initial graph containing both the high-level Event structure and fine-grained Plan–Observation candidates.

Importantly, the initial Matching Edges are treated as hypotheses rather than final groundings. Reliable matches provide candidate anchors for subsequent reasoning, while ambiguous or incorrect matches are handled by the following stages.

A.4.2 DIAGNOSE STAGE

The Construct stage favors broad candidate coverage and may therefore introduce noisy Plan–Observation Matching Edges. The Diagnose stage evaluates these existing relations and identifies groundings that are inconsistent with the surrounding narrative structure. It does not introduce new visual matches; instead, it updates the status and confidence of existing Matching Edges.

For each Plan–Observation edge, the model evaluates three complementary forms of consistency. **Temporal consistency** checks whether the candidate grounding is compatible with the ordering of neighboring Plans and their current matches. **Semantic consistency** evaluates whether the Plan semantics agree with the content represented by the matched Observation. **Event consistency** examines whether the match is coherent with the temporal span and narrative semantics of the Event containing that Observation.

The diagnosis is performed over the textual representations and metadata already stored in the graph, including Plan descriptions, Event summaries, Observation descriptions, and timestamps. Therefore, this stage does not require re-encoding the raw video.

Based on these signals, the model updates each Plan–Observation Matching Edge as reliable or uncertain and revises its confidence accordingly. Reliable edges are retained as **anchors**, whereas uncertain or conflicting edges are marked for further reasoning. The resulting anchor set provides stable contextual evidence for the Hypothesis stage, while unresolved Plans define the subset that requires additional inference.

A.4.3 HYPOTHESIS STAGE

The Hypothesis stage reasons about unresolved Plans at the Event level. Rather than directly searching again over all Observations, it establishes Plan–Event Relevance Edges that identify the narrative regions most likely to contain the missing evidence.

For each unresolved Plan, the model is provided with its textual representation, the current set of reliable anchors, and the available Event representations with their temporal information. It then reasons over three sources of structure: (1) semantic compatibility between the Plan and each Event; (2) temporal and contextual dependencies among neighboring Plans; and (3) the relative positions of already grounded anchor Plans and candidate Events.

The model selects a small set of plausible Events and creates the corresponding Plan–Event Relevance Edges. These edges do not directly determine the final grounding. Instead, they serve as structural hypotheses that restrict where subsequent visual verification is performed.

This stage is conducted over textual Plan and Event representations together with graph relations, rather than over the raw video. Consequently, long-range narrative reasoning is performed in the textual semantic space, while expensive multimodal processing is deferred to the final local verification stage.

A.4.4 VERIFY STAGE

The Verify stage converts the Event-level hypotheses into final task-specific predictions. For Movie Reasoning, this corresponds to refining subtitle-to-video grounding; the same stage can be instantiated with other task-specific outputs in different applications.

For each unresolved Plan, the model follows its Plan–Event Relevance Edges and retrieves only the Observations associated with the selected candidate Events. These Observations are then reintroduced as visual input together with the target Plan for localized multimodal verification. The model compares the candidate evidence, rejects inconsistent matches, and establishes or updates Plan–Observation Matching Edges accordingly.

Unlike the Construct stage, Verify does not process the complete video stream. Its multimodal input is restricted to the Observations contained in the hypothesized Events, substantially reducing the amount of visual evidence that must be reconsidered.

After verification, the updated Plan–Observation Matching Edges are used to produce the final task output. For Movie Reasoning, the selected Observation timestamps determine the final temporal grounding. The resulting graph therefore records both the final grounding relations and the intermediate Event-level hypotheses that led to them.

A.5 DATASET CURATION

Data Source. To construct a diverse and comprehensive benchmark, we collect movies in multiple languages, including English, Chinese, and Japanese. The dataset spans a wide range of genres, such as action, romance, and animation, and covers films released across different years. In addition, for each movie, we curate corresponding high-quality recap videos from online video platforms. Directly relying on human annotators to watch full-length movies for annotation is prohibitively expensive. Therefore, we design a semi-automatic data construction pipeline.

Narrative Query Construction. To evaluate high-level movie understanding rather than surface-level dialogue matching, we construct queries from recap narrations instead of original subtitles. We apply WhisperRadford et al. (2023) large-v3 to all recap videos to obtain ASR transcriptions, followed by inspection and lightweight post-processing. Each subtitle corresponds to a short temporal segment in the recap video, formulated as

$$S = \left\{ \left(s_j,\ t^{\text{recap}}_{j,\text{start}},\ t^{\text{recap}}_{j,\text{end}} \right) \right\}_{j=1}^{M}, \quad t^{\text{recap}}_{j,\text{start}},\ t^{\text{recap}}_{j,\text{end}} \in T^{\text{recap}}. \tag{5}$$

Here, s_j denotes the j-th subtitle sentence, $t^{\text{recap}}_{j,\text{start}}$ and $t^{\text{recap}}_{j,\text{end}}$ denote the start and end timestamps of the j-th subtitle in the recap video, and T^{recap} denotes the temporal range of the recap video.

Ground-Truth Annotation. The ground-truth correspondence between recap-subtitles and movie segments is automatically constructed through feature matching. For each source movie, we uniformly sample frames at a fixed rate r (2 FPS in our experiments). Let the sampled frame sequence be $F^{movie} = \{f_i\}_{i=1}^{N}$. Each frame is encoded by a pretrained visual encoder to obtain $z_i = \phi(f_i) \in \mathbb{R}^d$,

where $\phi(\cdot)$ denotes the feature extractor, Dinov3Siméoni et al. (2025). The entire movie is then represented as a temporally ordered feature bank: $Z^{movie} = [z_1, z_2, \ldots, z_N]$.

For each subtitle-aligned recap segment with temporal interval $[t^{recap}_{j,start}, t^{recap}_{j,end}]$, we sample frames from the corresponding recap clip and extract a feature sequence: $Q_j = [q_1, q_2, \ldots, q_L]$, where L is the number of sampled frames in the segment.

To preserve short-term temporal dynamics, we perform sequence-level matching rather than single-frame retrieval. Specifically, we compute the similarity between Q_j and every consecutive window of length L in the movie feature bank:

$$\text{score}(u) = \frac{1}{L} \sum_{k=1}^{L} q_k^\top z_{u+k-1}, \quad u^* = \arg\max_u \text{score}(u) \tag{6}$$

where u is the candidate start index. Finally, the automatically generated temporal annotation for Q_j is defined as $\hat{g}_j = [\tau_{u^*}, \tau_{u^*+L-1}]$, where τ_i denotes the timestamp of the i-th sampled movie frame.

Human Verification and Quality Control. To ensure the quality and rigor of the benchmark, we incorporated human review and supervision throughout the semi-automatic construction pipeline. For each movie, we first collected multiple candidate recap videos and manually selected the one with the highest overall quality, prioritizing comprehensive and coherent coverage of the original plot. During subtitle extraction and ground-truth construction, we performed sampled manual audits, achieving an estimated accuracy above 95%. This dataset contains 141 movies paired with their corresponding recap videos. The movie durations range from 44 to 260 minutes, while the recap videos range from 3 to 40 minutes. On average, each movie is associated with 471.60 subtitle segments.

B Additional Validation and Reproducibility Details

This section consolidates additional benchmark validation, evaluation, efficiency, and implementation details that complement the construction and four-stage descriptions above.

B.1 Release Scope and Benchmark Validation

Recap screening and subtitle localizability. Three annotators participated in the benchmark curation and validation process. Candidate recap videos were manually screened to ensure that they (i) primarily consist of clips from the corresponding movie without severe visual transformations that prevent reliable matching, (ii) cover the main narrative without omitting major events or introducing fabricated content, and (iii) are publicly accessible and can be processed under the final compliance protocol.

The 24,862 recap subtitle segments from the 50 evaluated movies were further categorized according to their temporal localizability. Among them, 15,971 segments were *explicitly localizable* to a unique movie event with relatively clear boundaries, 6,907 were *interval-localizable* but could correspond to a broader interval, several consecutive shots, or multiple visually similar candidates, and 2,084 were *not reliably localizable*, e.g., because they express abstract background information, character psychology, discontinuous events, or content without reliable visual correspondence. The last category is excluded as independent evaluation samples but may remain as contextual narration. Consequently, the evaluated benchmark contains 22,778 localized queries. Temporal intervals rather than single timestamps are used for valid ambiguous cases.

Verification of automatic matching. At least two annotators independently inspected each audited DINOv3 recap–movie match. Across 2,603 examined pairs drawn from all 50 evaluated movies, 2,487 were verified as correct, corresponding to an accuracy of 95.54% and a 95% Wilson confidence interval of 94.68–96.27%.

Candidate pool versus evaluated benchmark. The initial collection contains 141 movie–recap candidate pairs. Only the 50 movies that completed subtitle-localizability annotation, DINOv3-match verification, and final consistency checking constitute MRBench-Core and are used in all reported

Table 4: Temporal localization accuracy (%) by query type on the fixed partition of 22,778 evaluated queries. Columns report $\mathrm{Acc}@\delta$ for $\delta \in \{10, 30, 60, 120, 180, 300\}$ seconds.

	Local						Narrative / Context-dependent					
Method	10	30	60	120	180	300	10	30	60	120	180	300
Base (GPT-5.4)	31.55	49.98	61.02	63.20	66.11	66.29	12.52	21.19	26.15	31.37	33.05	35.69
Base (Gemini 3 Pro)	22.92	37.73	39.16	50.13	56.09	56.10	13.07	21.60	27.63	27.64	28.48	31.56
Base (Qwen3-VL-8B)	4.83	7.98	16.41	23.74	23.75	23.76	2.87	5.63	6.80	8.45	11.47	15.88
WorldMM (GPT)	46.60	65.85	71.61	71.62	71.64	71.65	10.61	19.82	26.35	32.06	34.57	35.71
WorldMM (Qwen3-VL-8B)	14.61	16.46	16.47	16.48	16.49	18.05	3.68	6.85	8.81	12.74	14.43	20.05
One-CLIP (Qwen3-VL-8B)	6.03	8.17	13.31	17.15	20.84	25.49	1.78	4.96	6.96	9.65	11.14	14.05
EdgeMemory (GPT-5.4)	31.52	56.13	65.09	73.68	76.98	77.01	15.85	28.02	38.44	47.05	50.87	56.21
EdgeMemory (Gemini 3 Pro)	11.66	19.00	26.83	33.52	33.53	42.98	19.34	33.61	40.41	41.61	42.20	55.76
EdgeMemory (Qwen3-VL-8B)	5.90	11.08	11.93	15.15	15.16	21.31	3.85	9.03	10.79	17.39	18.87	24.72

experiments. The remaining 91 pairs have undergone only partial automated processing and are excluded from the reported results until they complete the same validation protocol.

B.2 Local Grounding and Narrative-Dependent Queries

MRBench contains both locally resolvable queries and queries that require broader narrative context. We use the following binary operational criterion. A query is classified as *local* if one contiguous event provides sufficient grounding without requiring relations to other events. A query is classified as *narrative/context-dependent* if reliable localization requires cross-event evidence, for example to disambiguate repeated or visually similar events, resolve an abstract or distributed narrative description, or interpret references, causality, and implicit information.

Classification protocol and audit. For query-type annotation, GPT-5.6-sol receives only the marked target subtitle and the ordered subtitle transcript of the corresponding movie; it is not given video frames or the ground-truth segment. Each target is processed with one classification pass followed by one verification pass. To audit this categorization, we partitioned movies into five transcript-length strata spanning approximately 200–800 subtitles per movie and randomly selected one movie from each stratum. All 2,391 subtitles in these five movies were independently labeled by two annotators; disagreements were discussed and resolved by a third annotator. The automatic classification agreed with the adjudicated labels on 2,235 of 2,391 cases, corresponding to 93.48% accuracy.

The split reveals that narrative/context-dependent queries are substantially more difficult for direct and local-retrieval baselines. Under matched backbones, the relative gain of EdgeMemory is more pronounced on the narrative partition, while it does not uniformly outperform all alternatives on local queries. This analysis therefore supports a narrower interpretation of the task: MRBench is not exclusively a global-narrative benchmark, but a substantial portion of its difficulty arises from localization that cannot be resolved by isolated local visual matching.

B.3 Target-Duration Statistics

Table 5 reports the duration distribution of the ground-truth query-aligned intervals for all 22,778 evaluated queries. Most targets are temporally short: 88.39% are shorter than 20 seconds and 99.10% are shorter than 60 seconds. This distribution is useful for interpreting temporal-tolerance metrics and for distinguishing long-video search range from the duration of the localized evidence itself.

B.4 End-to-End Inference Cost

Per-call frame limits alone do not characterize the computational budget of a multi-stage long-video method. We therefore measure the complete inference pipeline under the same Qwen3-VL-8B backbone, evaluation split, and execution environment. The measurements include node construction and all subsequent edge updates. Table 6 reports average cost per movie for the five-movie efficiency audit.

Table 5: Duration distribution of ground-truth query-aligned intervals.

Duration	Count	Percentage (%)	Cumulative (%)
$[0, 5)$ s	11,745	51.56	51.56
$[5, 10)$ s	4,544	19.95	71.51
$[10, 20)$ s	3,845	16.88	88.39
$[20, 30)$ s	1,460	6.41	94.80
$[30, 60)$ s	979	4.30	99.10
$[60, 120)$ s	184	0.81	99.91
$[120, 300)$ s	21	0.09	100.00
≥ 300 s	0	0.00	100.00

Table 6: End-to-end inference cost under matched Qwen3-VL-8B settings. MM denotes multimodal calls.

Method	MM calls	Text calls	Frames	Input tokens	Output tokens	Min./movie
Qwen3-VL Direct	256.6	0	4,426.8	1,619,589	364,138	288.27
WorldMM	1,178.4	6,889.4	12,254.0	5,617,146	532,647	828.34
One-CLIP RAG	237.2	0	15,180.8	4,526,782	22,199	132.00
EdgeMemory	41.0	8.4	981.6	456,984.6	83,737	66.81

Under these matched settings, EdgeMemory uses fewer multimodal calls, cumulative frames, input tokens, and wall-clock time than the evaluated baselines. These numbers should be interpreted as a *relative end-to-end inference-efficiency comparison*, not as evidence that EdgeMemory is real-time or generally inexpensive: 66.81 minutes per movie remains a substantial absolute cost.

Qwen3-VL Direct protocol. For the same five-movie efficiency audit, Qwen3-VL Direct samples the movie at 0.1 FPS and forms non-overlapping windows of at most 50 frames (approximately 500 seconds). In Stage 1, each window is paired with all movie queries and produces query–frame candidates. Stage 2 performs one query-specific multimodal call to select among candidates returned from different windows. Every processed query in these five movies produced multiple candidates and therefore required this verification stage. The reported 256.6 multimodal calls is the average total number of calls per movie, not the number of calls per query.

B.5 External Generalization on EgoMemReason

To evaluate whether the memory mechanism transfers beyond movie temporal localization, we additionally evaluate EdgeMemory on EgoMemReason, which targets long-term memory and reasoning in egocentric videos. This benchmark differs from MRBench in video domain, task formulation, and evaluation dimensions, and therefore serves as a complementary generalization test rather than a direct reproduction of the MRBench localization setting.

EdgeMemory obtains an overall score of 36.9, compared with 29.6 for Qwen3-VL-8B and 30.6 for WorldMM. The largest absolute improvements are observed in Event Linking and the behavior-level spatial and activity dimensions, providing evidence that the structured-memory mechanism is useful beyond the movie-domain localization task.

Table 7: Generalization results on EGOMEMREASON (%).

Method	Entity Tracking	Entity Counting	Event Ordering	Event Linking	Behavior Spatial	Behavior Activity	Overall
Qwen3-VL-8B	35.0	28.0	23.0	21.0	40.0	42.0	29.6
WorldMM	32.0	44.0	21.0	21.0	34.0	36.0	30.6
EdgeMemory	31.0	42.0	25.0	40.0	48.0	45.0	36.9

B.6 NARRATIVE-HIERARCHY CONSTRUCTION AND METRIC SCOPE

The Beat–Sequence–Act hierarchy used for structural analysis is generated offline with GPT-5.4, fixed before model evaluation, and shared by all compared methods. Each movie is divided into overlapping 15-minute windows, with frames sampled at 0.1 FPS. The model first generates temporally bounded *Beats*. Adjacent Beats that contribute to the same local objective or event progression are grouped into *Sequences*, and consecutive Sequences are grouped into *Acts* according to the dominant objective, conflict state, and narrative stage. Strict temporal ordering and parent–child containment are enforced during construction.

We validate the generated hierarchy using automatic structural checks, human audits, and repeated-generation stability analysis. In particular, the audits examine the plausibility of temporal boundaries, semantic coherence within each unit, and consistency of parent–child relations across levels.

Importantly, Beat/Sequence/Act alignment is used only as a supplementary structural analysis. It measures whether predicted and ground-truth timestamps fall into the same region of a preconstructed narrative partition; it should not be interpreted as a direct measure of narrative understanding. The primary evaluation remains the hierarchy-independent temporal localization metric $\mathrm{Acc}@\delta$ computed from the verified temporal annotations.

B.7 OPERATIONAL SPECIFICATION OF EDGEMEMORY

The four-stage description in Sec. 4 specifies the reasoning flow. Here we provide the lower-level state representation and controller rules used in the reported implementation.

Node construction. Plan, Event, and Observation nodes are persistent after construction; memory evolution is implemented through edge addition, edge-state updates, and deprecation rather than rewriting previously constructed nodes. Adjacent recap subtitles are grouped into a Plan when they form a complete semantic unit. Observations are sampled at one frame every 10 seconds. Events are generated as atomic, temporally continuous units from 50-frame chunks with a one-frame overlap.

Typed relations and state. The implementation stores four principal code-level edge types:

Edge type	Direction	Role
`matched`	Observation → Plan	candidate/final grounding
`evidence_for`	Observation → Event	visual support for an Event
`related_to`	Event → Plan	established Event–Plan relation
`hypothesized_to`	Event → Plan	candidate Event for unresolved Plan

Each edge records its status, confidence label, update stage, and rationale. Edge status is one of `validated`, `tentative`, or `deprecated`. Confidence is categorical (`high`, `medium`, or `low`) and is not interpreted as a calibrated probability.

Construct and active-Plan control. Grounding uses no hidden cosine-similarity or scalar reliability threshold. A subtitle remains unmatched when the model returns the sentinel frame ID or fails to provide a valid rationale; any non-sentinel prediction must refer to an Observation supplied in the current frame stream. The controller keeps the earliest incomplete Plans in narrative order in an active pool: 10 Plans for API-model experiments and 6 Plans for the context-limited local-model

setting. Only unmatched subtitles from active Plans are exposed for assignment. Under the reported configuration, a Plan leaves the active pool after at least 50% of its subtitles have been matched. Its summary remains in context while the next incomplete Plan enters the pool. This 0.5 ratio controls pool progression only and is not an edge-confidence threshold.

Diagnose. Diagnose does not globally rematch subtitles. It conservatively flags only clear semantic, temporal, event-context, evidence, or rationale conflicts. Suspicious `matched` edges can be changed from validated to tentative so that they become eligible for subsequent hypothesis and verification, while their previous states remain available for provenance.

Hypothesis. For each unresolved Plan, Hypothesis receives the Plan state, existing temporal anchors, and a chronological table containing at most 200 Events. It returns one to five existing Event IDs per Plan, ranked according to semantic consistency, temporal proximity to existing anchors, and character/action/scene evidence. Invalid or duplicated Event IDs are removed. Before new ranked hypotheses are added, superseded hypotheses are deprecated rather than deleted.

Verify. Verify is invoked only for subtitles that remain unmatched or are flagged by Diagnose. It follows the ranked `hypothesized_to` and `evidence_for` edges to retrieve candidate Observations and orders them by Event rank and timestamp. Existing accepted subtitles in the same Plan are provided as fixed temporal anchors. Verification rematches only the affected subtitles while enforcing semantic relevance, within-Plan temporal order, and avoidance of collapsing multiple subtitles onto a single frame. Candidate Event groups are examined in rank order, and the first valid match is retained. If no valid candidate Event or supporting Observation is found, EdgeMemory does not fall back to an unconstrained global scan. Superseded matches are deprecated rather than deleted.

Common inference settings. Unless otherwise specified, the reported implementation uses temperature 0.2, 50-frame Construct chunks with a one-frame overlap, a Plan-completion ratio of 0.5, at most 200 Events supplied to Hypothesis, and at most five candidate Events per Plan.