LONGVIDGRAPH: PROGRESSIVE GRAPH MEMORY FOR RETRIEVAL-AUGMENTED OMNI-MODAL LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video understanding with omni-modal models remains challenging due to limited context budgets and long-range temporal dependencies. Existing methods rely on dense frame sampling or simple context accumulation, which fail to maintain coherent information over long durations. Furthermore, existing memory mechanisms mainly rely on unstructured captions, limiting cross-segment memory consolidation and retrieval. We propose LongVidGraph, a retrieval-augmented framework with a progressive graph memory. The model incrementally constructs a structured graph of entities, actions, and dialogues from audio-visual signals in long videos. This graph serves as a compact and temporally consistent memory. To enable the model to learn this incremental and structured representation, we build a large-scale clip-graph supervision pipeline that produces aligned video-graph training data at scale. We fine-tune Qwen2.5-Omni on this data for progressive graph construction. At inference time, we perform query-conditioned retrieval over the graph memory to obtain relevant structured evidence. The retrieved evidence is used to support downstream understanding tasks. Experiments on multiple long-video benchmarks show consistent improvements over strong open-source baselines, with more significant gains on long-duration videos. The results demonstrate the effectiveness of progressive graph memory and retrieval for long-video understanding with omni models.

1 INTRODUCTION

Long-video understanding has become an important research direction in video-language modeling, requiring models to comprehend visual, textual, and acoustic information over extended temporal contexts. Recent omni-modal large language models provide a promising foundation for this task by jointly modeling video, audio, and language within a unified framework. Most video understanding models represent videos by sampling frames at a predefined rate within a fixed context window. As video duration increases, dense sampling quickly exceeds the model's context capacity, forcing existing methods to adopt sparser sampling. Although this improves temporal coverage, it inevitably misses fine-grained information, such as transient actions, state transitions, and local interactions, limiting the model's ability to capture detailed event evolution in long videos.

To address the challenge of efficiently accessing information from long contexts, retrieval-augmented generation (RAG) and memory-based approaches have been widely explored in language and multimodal systems. This paradigm provides a promising direction for long-video understanding by transforming temporally distributed observations into persistent memory representations. However, existing video memory mechanisms mainly rely on caption-based summaries at different temporal granularities. Such unstructured representations lack explicit relationships among entities, events, and temporal dependencies, making cross-segment memory consolidation difficult and introducing irrelevant or redundant information during retrieval. Therefore, a structured memory representation is needed to support effective long-range organization and retrieval for omni-modal long-video understanding.

In this work, we introduce LongVidGraph, a progressive-graph-based framework that improves the long-video understanding capability of omni-modal models through progressive graph construction

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.

and question-guided evidence retrieval. We first build an automated data construction pipeline to provide supervision for progressive video graph generation. Specifically, we collect and temporally align the audio and visual content of more than 400 long videos, use large multimodal models to extract the atomic information required by the graph, and employ a large language model to organize these observations into a coherent graph structure. This pipeline yields approximately 15K clip-graph training pairs for learning progressive graph construction.

We then fine-tune Qwen2.5-Omni on the constructed data, enabling it to process a long video clip by clip and generate a delta graph for each newly observed segment. The delta graphs are progressively merged into a complete structured representation of the video. Based on the resulting graph, we further design a retrieval-based question answering pipeline that decomposes the graph into fine-grained retrieval units. For each question, the pipeline retrieves the most relevant graph elements and uses their associated timestamps to extract a small set of supporting frames from the original video. The retrieved textual and visual evidence is then jointly incorporated into downstream video question answering, providing compact yet informative context for answer prediction. Experiments across multiple video understanding benchmarks demonstrate that our framework consistently improves question answering performance and achieves particularly significant gains on long-video benchmarks, validating the effectiveness of structured progressive memory for preserving and retrieving fine-grained temporal information.

Our main contributions are summarized as follows:

- **Progressive graph data construction.** We develop an automatic data construction pipeline that converts video clips into structured atomic information and progressively merges them into temporally coherent video graphs. The resulting dataset provides multimodal supervision for learning progressive video graph generation.
- **Omni-modal progressive memory learning.** We fine-tune Qwen2.5-Omni on the constructed dataset, equipping the model with the ability to generate structured delta graphs from sequential video clips and maintain information across extended video durations.
- **Graph-based evidence retrieval.** We propose a question-guided retrieval pipeline that decomposes a complete video graph into fine-grained semantic elements, retrieves question-relevant subgraphs, and extracts timestamp-grounded visual evidence for downstream question answering.
- **Extensive benchmark evaluation.** We evaluate our method on multiple video understanding benchmarks with different video lengths and reasoning requirements. The results demonstrate consistent improvements over existing approaches, with especially pronounced advantages on long-video question answering.

2 Related Work

2.1 Long Video Understanding

Recent multimodal large language models (MLLM) have extended video understanding from short clips to longer temporal contexts. Several studies seek to represent long videos through textual caption (Wang et al., 2024; Wei et al., 2025; Xiong et al., 2024) at different temporal granularities. Video ReCap recursively generates clip-, segment-, and video-level captions, combining detailed local events with high-level summaries (Islam et al., 2024). Chapter-Llama (Ventura et al., 2025) integrates ASR transcripts with captions of selected frames and uses an LLM to divide hour-long videos into semantically coherent chapters with corresponding titles. ARC-Chapter (Pu et al., 2025) further constructs temporally grounded and hierarchical chapter annotations, ranging from concise titles to detailed summaries, by integrating speech transcripts, scene text, and visual captions.

Another line of work employs agentic (Long et al., 2025; Yang et al., 2025) or coarse-to-fine frameworks (He et al., 2025; Ye et al., 2025) to inspect long videos at finer temporal granularity through iterative tool use and adaptive zoom-in. LVAgent coordinates multiple multimodal agents through iterative discussion and refinement (Chen et al., 2025), while LongVideoAgent (Liu et al., 2026) assigns temporal grounding and local visual perception to specialized agents under the planning of a master model. ZoomV (Pan et al.) follows a coarse-to-fine strategy that progressively narrows the temporal search range, and VideoZoomer (Ding et al., 2025) learns to dynamically determine when

and where to inspect the video at a finer temporal resolution. These methods allocate more perception and reasoning effort to potentially relevant video regions. Nevertheless, caption-based approaches may lack structured relationships for effective memory consolidation and retrieval, whereas agentic approaches often rely on complex, model-specific workflows and strong proprietary models, which may limit their reproducibility and generalizability. Our work instead learns an end-to-end incremental graph construction model to preserve fine-grained video information.

2.2 Graph-based Memory

Graph representations have long been used to organize entities and their relations in videos. MovieGraphs (Vicol et al., 2018) introduces human-centric graphs for movie clips, describing characters, attributes, interactions, and social relations within individual scenes. Subsequent studies extend such static character graphs along the temporal dimension. Dynamic Character Graph (Kulshreshtha & Guha, 2020) constructs evolving character interaction networks from full-length movies, while MGRG (Xu et al., 2025a) aggregates relation graphs at multiple temporal granularities for social relation reasoning. More recently, mtKG-LLM (Li & Xu, 2024) divides long videos into scenes, constructs a multimodal spatial knowledge graph for each scene, and continuously updates a temporal knowledge graph to support long-term social relation recognition.

Beyond character-centric analysis, video scene graph methods represent videos as sequences of object-relation triplets and have been applied to temporal reasoning (Cong et al., 2021; Geng et al., 2020; Cherian et al., 2022; Chu et al., 2025). Graph Prompts (Li et al., 2025) further employs pretrained Video Scene Graph Generation models to construct video graphs and transforms their structured information into prompts for downstream tasks. More recent studies explore graph-based representations for long egocentric videos (Rege et al., 2026). VideoMindPalace (Huang et al., 2026) organizes important moments into an environment-grounded hierarchical semantic graph, capturing information from fine-grained object interactions to higher-level spatial regions. These studies demonstrate the potential of graphs for preserving relational and temporal structures in videos. However, existing approaches mainly use static graph representations or focus on specific tasks such as social relation recognition and short-video reasoning. In contrast, we learn an incremental graph memory from multimodal long-video observations for general long-video question answering.

3 Method

3.1 Progressive Video Graph Representation

We represent a long video as a temporally evolving multimodal graph that is updated over consecutive short video segments. Rather than reconstructing the complete video structure at each step, the model generates a *graph snapshot* for the currently observed interval. Each snapshot records newly observed entity information, actions, and dialogues, together with their temporal boundaries. The snapshots are subsequently accumulated to form a structured representation of the entire video.

Formally, given a divided video segments $V_{t_{t-1}}^{T}$, the model produces a progressive graph update

$$\Delta\mathcal{G}_t = (\mathcal{E}_t, \mathcal{A}_t, \mathcal{D}_t)\,, \tag{1}$$

where $\mathcal{E}_t$, $\mathcal{A}_t$, and $\mathcal{D}_t$ denote the entity, action, and dialogue information observed within the temporal interval $[\tau_t^s, \tau_t^e]$, respectively. The accumulated video graph is updated as

$$\mathcal{G}_t = \mathrm{Merge}\left(\mathcal{G}_{t-1}, \Delta\mathcal{G}_t\right), \tag{2}$$

where the merge operation preserves existing identities and appends newly observed information.

Entities. Each entity has a persistent identifier across video segments. For each entity, the graph maintains two temporally grounded streams: *appearance* (visual characteristics over time) and *state* (condition, role, location, emotion, or other dynamic properties). This separation distinguishes stable visual traits from evolving semantics and explicitly tracks entity changes over time.

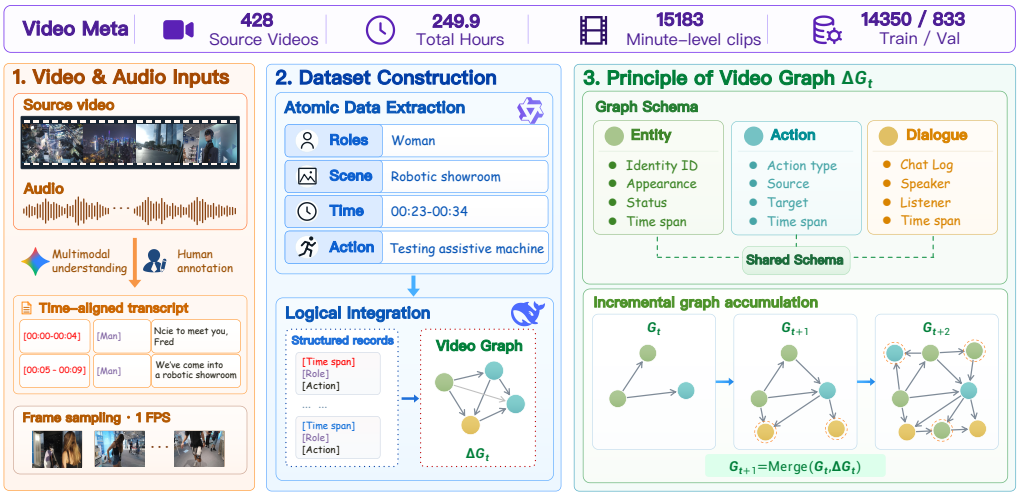

Figure 1: **Overview of the LongVidGraph dataset construction**. It presents the dataset statistics, the incremental video graph representation consisting of entity, action, and dialogue nodes, and the construction pipeline for extracting atomic evidence and integrating it into structured video graphs.

Actions. Actions represent temporally localized interactions with one source entity and zero or more target entities. Each action is defined by a semantic type, participating entities, and start/end timestamps. This supports both individual behaviors and multi-entity interactions while preserving direction via the source-target structure.

Dialogues. Dialogue entries capture speech-related interactions derived from aligned audio-visual observations. Each entry records the speaker, potential listeners, temporal boundaries, and a concise semantic summary of the utterance. Instead of storing raw transcripts alone, the graph retains communicative content and participant structure, enabling retrieval of both the spoken content and the involved parties.

Overall, the graph serves as a progressively built, persistent, fine-grained memory, where entity nodes endure across segments, action and dialogue relations accumulate over time, and appearance/state streams continuously update to reflect evolving attributes. Unlike isolated clip captions, this representation maintains cross-segment identity and temporal continuity, providing a compact yet expressive memory foundation for long-video reasoning.

3.2 Dataset Construction

3.2.1 Video Meta

Our source corpus consists of long-form videos. After deduplication by source identity, we obtain 428 source videos in total, with 400 videos for training and 28 videos for validation.

We segment each source video using a fixed 1-minute temporal window (60,000 ms), producing 16,820 minute-level clips in the initial slicing stage (the last window of a video may be shorter than 60 seconds). After downstream filtering and sample packaging, the final dataset contains 15,183 clips, including 14,350 training samples and 833 validation samples.

Measured on the final retained samples, the total clip duration is 236.22 hours for training and 13.68 hours for validation, i.e., 249.90 hours in total.

3.2.2 Three-Stage Graph Generation Pipeline

As illustrated in Figure 1, we construct minute-level graph supervision through a three-stage data generation pipeline that progressively converts raw long videos into progressive graph targets.

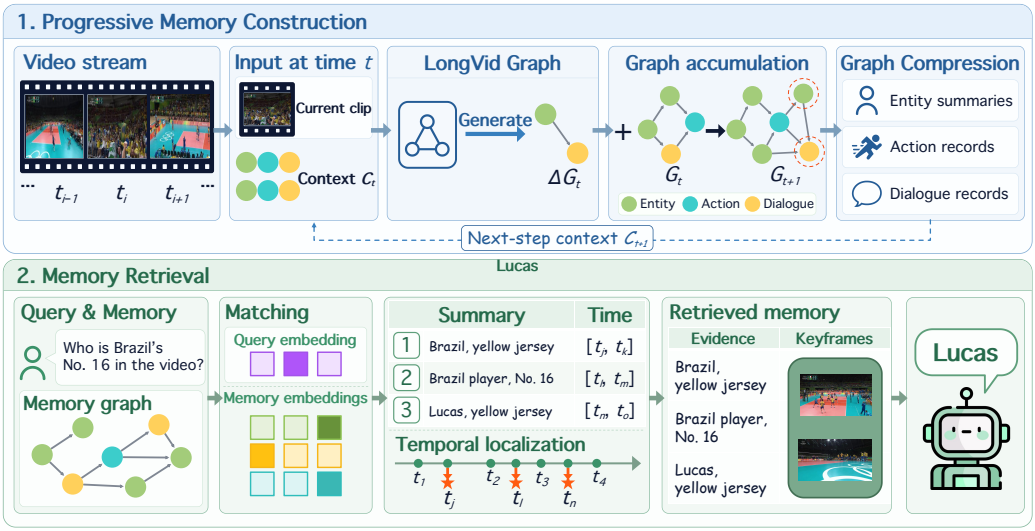

Figure 2: Overview of LongVidGraph with progressive memory construction and memory retrieval.

Stage I: Temporally aligned speech transcription. We first obtain ASR transcripts from each video and align every utterance with its temporal span and speaker identity. The resulting timestamped dialogue segments provide reliable linguistic cues for subsequent multimodal analysis.

Stage II: Audio-visual evidence extraction. Each video is divided into one-minute clips. We use Qwen3.5-397B-A17B (Qwen Team, 2026) to processes every clip together with the aligned dialogue segments within the same temporal window, extracting the atomic evidence required for graph construction, including entity observations, actions, and conversational events. At this stage, the model focuses on identifying information directly grounded in the audio-visual content rather than organizing it into a complete graph.

Stage III: Progressive graph consolidation. Finally, DeepSeek-R1 (Guo et al., 2025) is conditioned on both the current 1-minute clip evidence and the accumulated graph context G_{t-1}, and predicts only the minute-level delta graph ΔG_t. The updated graph is then obtained by a deterministic rule-based merge, $G_t = \mathrm{Merge}(G_{t-1}, \Delta G_t)$. Concretely, the merge keeps historical facts intact (append-only update), merges entities by identity, appends newly predicted appearances/states/actions/dialogues, removes exact duplicates by predefined keys, and finally reorders events by time. This design enforces stable progressive graph growth across consecutive clips while preventing repetitive regeneration of previously known content.

This decomposition separates speech alignment, multimodal evidence extraction, and cross-clip logical consolidation, allowing each model to focus on a well-defined objective and improving the temporal consistency of the generated training data.

3.3 Graph-Conditioned QA Pipeline

As shown in Figure 2, LongVidGraph first constructs a progressive graph memory from long videos and then retrieves relevant evidence from this memory for downstream question answering. Building on the clip-graph supervision described above, we fine-tune Qwen2.5-Omni (Xu et al., 2025b) to progressively extract structured information from minute-level video clips.

(1) Progressive memory construction. We process each long video into minute-level clips and update the graph memory sequentially. At minute t, the model takes the current clip V_t together with a compact representation of the historical graph context C_t, and predicts newly observed graph information ΔG_t. Formally,

$$\Delta G_t = f(V_t, C_t), \qquad C_t = \mathcal{C}(G_{t-1}),$$

where G_{t-1} denotes the accumulated graph before processing the current clip and $\mathcal{C}(\cdot)$ is a structured graph compression function.

Instead of serializing the full historical graph, we retain only compact information required for subsequent extraction. Specifically, C_t is represented as line-based records containing: (i) entity summaries with identity, latest state, and appearance history; (ii) recent action records; and (iii) recent dialogue records within a bounded temporal window (3 minutes). This substantially reduces context-token overhead while preserving the historical information needed for progressive graph construction.

The predicted delta graph ΔG_t is then incorporated into the running memory through a deterministic merge operator,

$$G_t = \mathcal{M}(G_{t-1}, \Delta G_t).$$

By iterating over all clips, we obtain a video-level graph that progressively accumulates entities, actions, dialogues, and their relations, serving as the structured memory for subsequent retrieval and question answering.

(2) Question-conditioned memory retrieval. Given a question q, we linearize graph elements into a retrieval corpus and perform dense semantic retrieval with cosine similarity. In practice, we use a sentence-embedding retriever and keep the top-K evidence items with $K = 15$. For each retrieved item, we take the midpoint timestamp $(s_i + e_i)/2$, cluster nearby timestamps with a 3-second merge threshold, and retain at most 16 key moments as visual evidence anchors. These anchors are mapped to video frames, while the retrieved graph items are temporally reordered to form a question-conditioned compact graph summary for downstream answer generation.

4 Experiment

4.1 Settings

We fine-tune Qwen2.5-Omni via LoRA (Hu et al., 2021) adaptation (r=8, α=32, applied to all linear layers) using the ms-swift (Zhao et al., 2024) framework. Training is distributed across 16 GPUs (2 nodes $\times$ 8). We train for 3 epochs at a learning rate of 1×10^{-4} with an effective batch size of 128. For video inputs, we sample at 1 FPS with up to 60 frames and 1 million pixels per video, with audio modality enabled.

We evaluate LongVidGraph on four widely adopted long-video understanding benchmarks: Video-MME (Fu et al., 2025; 2026), LVBench (Wang et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2025), covering diverse long-form video understanding and question-answering scenarios. To ensure all results reflect generalization to unseen videos, the fine-tuning datas are completely independent of the evaluation datasets. In addition to Qwen2.5-Omni, which serves as our primary baseline, we compare against two open-source omni-modal models, Qwen3-Omni (Xu et al., 2025c) and MiniCPM-o 2.6 (Yao et al., 2024), as well as two proprietary models, GPT-4o (OpenAI, 2024) and Gemini 1.5 Pro (Gemini Team, 2024). For all open-source baselines, we uniformly sample videos at 1 FPS, with a maximum of 64 frames per video.

Table 1: Comparison on Video-MME and MLVU.

Source	Model	Video-MME				MLVU			
		0-5 min	5-30 min	30 min+	Overall	3-10 min	10-20 min	20 min+	Overall
Closed	Gemini 1.5 Pro	81.70%	74.30%	67.40%	75.00%	–	–	–	–
	GPT-4o	80.00%	70.30%	65.30%	71.90%	–	–	–	64.60%
Open	Qwen3-Omni	70.78%	59.22%	54.78%	61.59%	62.96%	70.62%	53.62%	63.62%
	MiniCPM-o 2.6	68.11%	57.00%	48.67%	57.93%	61.23%	59.16%	55.00%	60.36%
	Qwen2.5-Omni (Baseline)	65.11%	53.89%	46.44%	55.15%	56.84%	63.28%	44.93%	57.13%
	LongVidGraph (Ours)	65.22%	56.67%	52.44%	58.11%	62.43%	66.38%	55.07%	62.60%

4.2 Main Results

As shown in Table 1, LongVidGraph consistently improves upon the Qwen2.5-Omni baseline on both Video-MME and MLVU, with the gains becoming more pronounced as video duration increases. On Video-MME, LongVidGraph improves the overall score from 55.15% to 58.11%, while the gain on the 30 min+ subset reaches 6 points (46.44% $\rightarrow$ 52.44%), compared with only a marginal difference on 0–5 min videos. A similar trend is observed on MLVU, where the overall performance

Table 2: Comparison on LongVideoBench and LVBench.

Source	Model	LongVideoBench				LVBench			
		0-10 min	10-30 min	30-60 min	Overall	30-60 min	60-90 min	90 min+	Overall
Closed	Gemini 1.5 Pro	–	–	–	64.00%	–	–	–	33.10%
	GPT-4o	–	–	–	66.70%	–	–	–	48.90%
Open	Qwen3-Omni	67.10%	51.09%	58.33%	60.88%	37.31%	38.44%	40.34%	38.35%
	MiniCPM-o 2.6	59.69%	44.13%	53.70%	53.85%	33.20%	38.23%	31.37%	34.28%
	Qwen2.5-Omni (Baseline)	57.22%	42.17%	50.93%	51.53%	35.80%	34.77%	35.85%	35.51%
	LongVidGraph (Ours)	57.61%	46.09%	58.33%	53.70%	40.74%	39.09%	39.78%	40.03%

Table 3: Ablation study on the contributions of textual graph information and keyframes under different QA models. The best result in each QA-model group is highlighted in **bold**.

Model	Text	Keyframe	Video-MME				LVBench			
			0-5 min	5-30 min	30 min+	Overall	30-60 min	60-90 min	90 min+	Overall
Qwen2.5-Omni	✓		54.44%	49.78%	47.33%	50.52%	34.16%	30.45%	29.69%	32.02%
		✓	58.33%	52.56%	51.67%	54.19%	**39.64%**	36.72%	35.01%	37.70%
	✓	✓	**64.78%**	**55.22%**	**53.56%**	**57.85%**	39.51%	**39.09%**	**38.38%**	**39.12%**
LongVidGraph	✓		55.22%	50.22%	48.33%	51.26%	33.74%	32.40%	31.65%	32.86%
		✓	61.11%	52.89%	52.11%	55.37%	40.19%	38.66%	36.97%	38.99%
	✓	✓	**65.22%**	**56.67%**	**52.44%**	**58.11%**	**40.74%**	**39.09%**	**39.78%**	**40.03%**

increases from 57.13% to 62.60%, and the improvement is particularly substantial for 20 min+ videos, from 44.93% to 55.07% (+10.14 points). These results suggest that the proposed structured graph memory is especially effective in mitigating information loss over extended temporal contexts. Compared with the similarly sized MiniCPM-o 2.6, LongVidGraph also achieves competitive or better performance; for example, it surpasses MiniCPM-o 2.6 on the Video-MME 30 min+ subset by 3.77 points (52.44% vs. 48.67%) and slightly improves the overall score (58.11% vs. 57.93%). Despite using a smaller backbone than Qwen3-Omni, LongVidGraph remains competitive, particularly on long-duration videos: on MLVU 20 min+, it even exceeds larger Qwen3-Omni (55.07% vs. 53.62%), while its MLVU overall score remains within 1.02 points (62.60% vs. 63.62%).

Table 4: Ablation study of the retrieval graph on LVBench.

Model	Retrieved Memory	LVBench			
		Short	Med.	Long	Overall
Qwen3-Omni		37.31%	38.44%	40.34%	38.35%
	✓	40.74%	42.98%	40.90%	41.45%
MiniCPM-o 2.6		33.20%	38.23%	31.37%	34.28%
	✓	40.19%	38.66%	42.02%	40.15%
Qwen2.5-Omni		35.80%	34.77%	35.85%	35.51%
	✓	39.51%	39.09%	38.38%	39.12%

Table 2 further highlights the competitiveness of LongVidGraph on longer and more challenging videos. LongVidGraph performs particularly well on the long-duration subsets. On LongVideoBench, it matches Qwen3-Omni on the 30-60 min subset (58.33%), while on LVBench it achieves a higher overall score of 40.03% compared with 38.35%, and remains close on the 90 min+ subset (39.78% vs. 40.34%). Compared with the similarly sized MiniCPM-o 2.6, LongVidGraph shows clearer advantages on long-duration videos, outperforming it by 4.63 points on the 30-60 min subset of LongVideoBench (58.33% vs. 53.70%) and by 8.41 points on the 90 min+ subset of LVBench (39.78% vs. 31.37%). These gains are also consistently observed over the Qwen2.5-Omni baseline. On LongVideoBench, our method improves the overall score from 51.53% to 53.70%, with the gain increasing from 3.92 points on 10-30 min videos to 7.40 points on 30-60 min videos. On LVBench, it improves Qwen2.5-Omni across all duration ranges, with gains of 4.94, 4.32, and 3.93 points on the 30-60 min, 60-90 min, and 90 min+ subsets, respectively, resulting in an overall improvement from 35.51% to 40.03%. These results further indicate that the proposed graph-based memory is particularly beneficial as video duration increases.

4.3 More Results

Case Study. As shown in Fig. 3, our retrieval strategy benefits both cumulative-evidence and snapshot-evidence questions. For cumulative-evidence questions, the key is to collect consistent observations across the video rather than rely on a single frame. In the first example, the question asks for the man's primary exercise type. Our method retrieves several clips showing leg and lower-

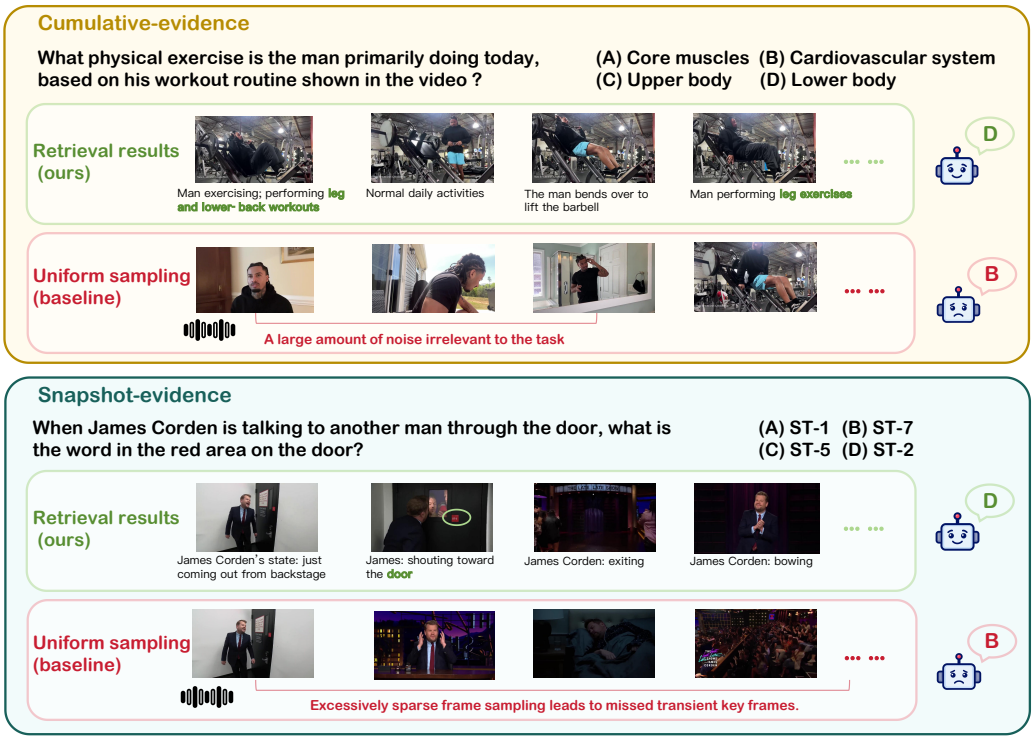

Figure 3: Qualitative results on cumulative-evidence and snapshot-evidence questions. Our method consistently retrieves more relevant visual evidence for both temporally accumulated reasoning and sparse key-frame recognition, leading to more reliable answers than uniform sampling.

back workouts, which together support the correct answer ”Lower body”. In contrast, uniform sampling contains many irrelevant daily-activity frames and leads to an incorrect prediction.

For snapshot-evidence questions, the challenge is different: the answer may appear only in a brief moment. In the second example, the question asks for the word in the red area on the door. Our method retrieves the transient frame containing the required visual cue and produces the correct answer, while sparse uniform sampling misses this key moment. These two cases show that our retrieval mechanism can both aggregate distributed evidence and preserve sparse but decisive observations in long videos.

Effect of Graph Construction Training. A further comparison between Qwen2.5-Omni and LongVidGraph reveals an additional benefit of clip-graph supervision. In Table 3 under the same setting, LongVidGraph consistently achieves better performance, although it is trained only with clip-graph construction data rather than QA supervision. For example, with text and keyframes, LongVidGraph improves the overall score from 57.85% to 58.11% on Video-MME and from 39.12% to 40.03% on LVBench. The improvement is more evident with keyframe-only evidence, where the overall scores increase from 54.19% to 55.37% on Video-MME and from 37.70% to 38.99% on LVBench. These results suggest that learning to construct structured graph memory also implicitly improves the model's ability to reason over retrieved evidence for question answering.

4.4 Ablations

Retrieved Modalities. We conduct the modality ablation on Qwen2.5-Omni and LongVidGraph over Video-MME and LVBench in Table 3. Combining textual and visual retrieval consistently achieves the best performance, demonstrating that structured semantic representations and visual evidence provide complementary information for long-video understanding. With LongVidGraph, text-only retrieval achieves 51.26% and 32.86% overall on Video-MME and LVBench, compared with 55.37% and 38.99% using keyframes alone. Combining text and keyframes yields the best overall performance, reaching 58.11% and 40.03%, respectively. For relatively shorter videos, visual evidence alone can already be competitive; on the 30-60 min subset of LVBench, Qwen2.5-Omni with

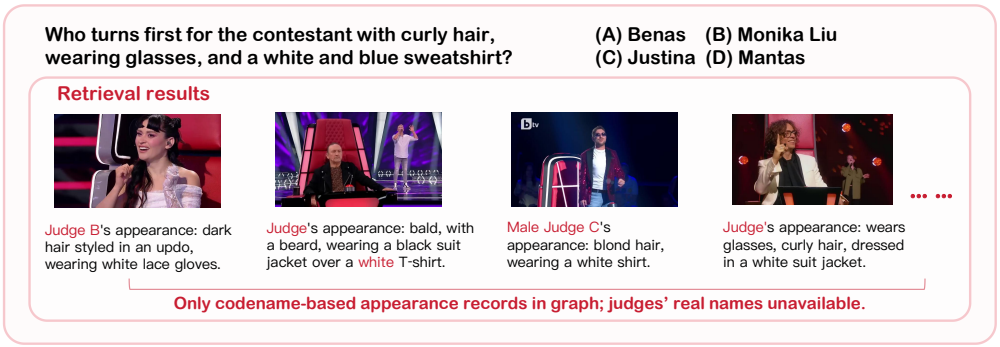

Figure 4: A failure case of LongVidGraph with incomplete graph memory.

keyframes achieves 39.64%, slightly higher than the 39.51% obtained with both modalities. In contrast, the benefit of multimodal retrieval becomes clearer on longer videos: on the 90 min+ subset, combining text and keyframes improves performance from 35.01% to 38.38% with Qwen2.5-Omni and from 36.97% to 39.78% with LongVidGraph.

Retrieved Memory. We further evaluate whether the evidence retrieved from LongVidGraph can benefit different QA models. We inject the retrieved graph evidence into three open-source models, and conduct experiments on LVBench, which contains a large proportion of long videos. As shown in Table 4, retrieval augmentation consistently improves the overall performance of all three models, from 38.35% to 41.45% for Qwen3-Omni, from 34.28% to 40.15% for MiniCPM-o 2.6, and from 35.51% to 39.12% for Qwen2.5-Omni. The gains are particularly pronounced on long videos. On the longest subset, MiniCPM-o 2.6 improves from 31.37% to 42.02%, while Qwen2.5-Omni increases from 35.85% to 38.38%. These results demonstrate that LongVidGraph provides model-agnostic retrieval evidence and is especially beneficial for long-duration video understanding.

4.5 Discussion

Despite the advantages, LongVidGraph still has limitations. First, graph quality depends on the capability of the underlying omni-modal model. Extraction errors can propagate into the memory, including inaccurate timestamps and incorrect visual attributes. In Figure 4, for example, a brown shirt is incorrectly recorded as white. Second, graph construction is currently open-loop. Extracted information is written into the graph without later correction or feedback. This is especially problematic for cross-clip identity association. In the shown case, retrieval finds appearance records that match the queried visual characteristics, but these records are only associated with temporary role-based identifiers. The graph does not contain the person's name required by the question, so the retrieved evidence is insufficient to answer it.

These limitations suggest two directions for future improvement. First, integrating external tools such as visual grounding (Liu et al., 2024), tracking (Zhang et al., 2022) and entity matching (Ren et al., 2023) could improve the reliability of graph construction, especially for cross-clip identity association. Second, future systems could move graph construction toward adaptive memory updating, where incomplete graph entries are refined based on retrieval results and downstream feedback.

5 Conclusion

In this work, we introduce LongVidGraph, a retrieval-augmented framework for long-video understanding with omni-modal models. LongVidGraph progressively constructs a structured graph memory from audio-visual signals over time. We further build a large-scale clip-graph supervision pipeline to train the model for incremental graph construction, and perform query-conditioned retrieval over the accumulated memory to provide relevant evidence for downstream QA tasks. Extensive experiments on Video-MME, LVBench, LongVideoBench, and MLVU show consistent improvements over strong baselines, with larger gains on longer videos. These results validate the effectiveness of progressive structured memory and retrieval for long-video understanding, and suggest a promising direction for extending omni-modal models beyond fixed-context processing.

AI USE STATEMENT

Generative AI tools were used for dataset construction, prompt drafting, code and figure generation, data filtering, and language editing. Qwen3.5-397B-A17B and DeepSeek-R1 were used to generate structured graph annotations as described in Section 3.2. All outputs used in this work were reviewed and validated by the authors, who take full responsibility for the final content and claims.

ETHICS STATEMENT

This work does not involve direct interaction with human subjects. The source videos are openly available but may contain identifiable individuals and speech. We review data provenance and redistribution conditions during curation and acknowledge potential risks related to privacy, licensing, dataset bias, and annotation errors.

REFERENCES

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, 2025.

Anoop Cherian, Chiori Hori, Tim K. Marks, and Jonathan Le Roux. (2.5+1)d spatio-temporal scene graphs for video question answering, 2022. URL https://arxiv.org/abs/2202.09277.

Meng Chu, Yicong Li, and Tat-Seng Chua. Understanding long videos via llm-powered entity relation graphs, 2025. URL https://arxiv.org/abs/2501.15953.

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation, 2021. URL https://arxiv.org/abs/2107.12309.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding, 2026. URL https://arxiv.org/abs/2604.05015.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Shijie Geng, Peng Gao, Chiori Hori, Jonathan Le Roux, and Anoop Cherian. Spatio-temporal scene graphs for video dialog. *ArXiv*, abs/2007.03848, 2020. URL https://api.semanticscholar.org/CorpusID:220404603.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean

Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Yiqiu Ren, Licheng Yu, Ning Zhang, Yong Jae Lee, and Miao Liu. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms, 2026. URL https://arxiv.org/abs/2501.04336.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18198–18208, 2024.

Prakhar Kulshreshtha and Tanaya Guha. Dynamic character graph via online face clustering for movie analysis, 2020. URL https://arxiv.org/abs/2007.14913.

DaiFeng Li and Fan Xu. Synergizing knowledge graphs with large language models: A comprehensive review and future prospects, 2024. URL https://arxiv.org/abs/2407.18470.

Yiming Li, Xiaoshan Yang, Bing-Kun Bao, and Changsheng Xu. Graph prompts: Adapting video graph for video question answering. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 1485–1493. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/166. URL https://doi.org/10.24963/ijcai.2025/166. Main Track.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. Longvideoagent: Multi-agent reasoning with long videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40404–40416, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Junwen Pan, Yuan Zhang, Rui Zhang, Xin Wan, Qizhe Zhang, Ming Lu, Shanghang Zhang, and Qi She. Zoomv: Temporal zoom-in for efficient long video understanding.

Junfu Pu, Teng Wang, Yixiao Ge, Yuying Ge, Chen Li, and Ying Shan. Arc-chapter: Structuring hour-long videos into navigable chapters and hierarchical summaries. *arXiv preprint arXiv:2511.14349*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL 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. Agentic very long video understanding, 2026. URL https://arxiv.org/abs/2601.18157.

Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, and Xiaokang Yang. Facial geometric detail recovery via implicit representation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Chapter-llama: Efficient chaptering in hour-long videos with llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18947–18958, 2025.

Paul Vicol, Makarand Tapaswi, Lluís Castrejón, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8581–8590, 2018. doi: 10.1109/CVPR.2018.00895.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2025. URL https://arxiv.org/abs/2406.08035.

Yu Wang, Zeyuan Zhang, Julian McAuley, and Zexue He. Lvchat: Facilitating long video comprehension. *arXiv preprint arXiv:2402.12079*, 2024.

Hongchen Wei, Zhihong Tan, Yaosi Hu, Chang Wen Chen, and Zhenzhong Chen. Longcaptioning: Unlocking the power of long video caption generation in large multimodal models. *arXiv preprint arXiv:2502.15393*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. *Advances in Neural Information Processing Systems*, 37:16623–16644, 2024.

Cong Xu, Feiyu Chen, Qi Jia, Yihua Wang, Liang Jin, Yunji Li, Yaqian Zhao, and Changming Zhao. A multi-granularity relation graph aggregation framework with multimodal clues for social relation reasoning. *IEEE Transactions on Multimedia*, 27:4961–4970, 2025a. doi: 10.1109/TMM.2025.3543054.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025b. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025c. URL https://arxiv.org/abs/2509.17765.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xinlin Zhuang, Yifan Lu, Xiang An, Ming Hu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, et al. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8579–8591, 2025.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022. URL https://arxiv.org/abs/2110.06864.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding, 2025. URL https://arxiv.org/abs/2406.04264.