Because Flat Retrieval Is Not Enough: VideoGraph for Long Video Question Answering

Dany Chahine [1] **Ali Chehab** [2] **Ammar Mohanna** [2]

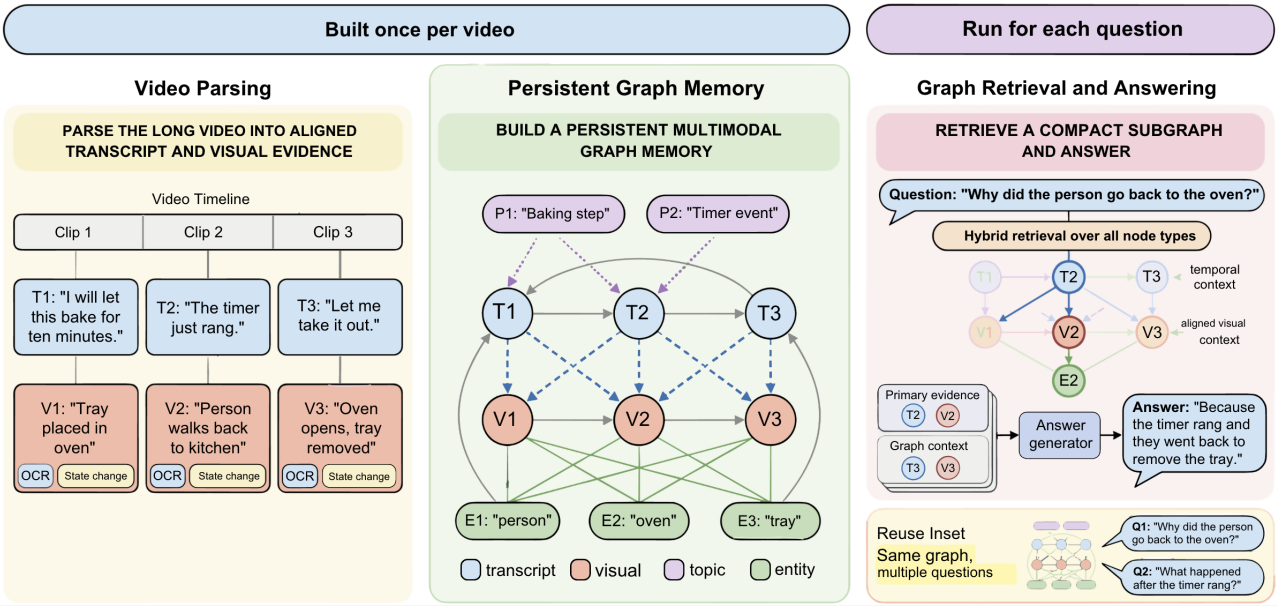

Figure 1. Overview of VideoGraph. A long video is parsed into aligned transcript segments and visual clips, which are organized into a persistent multimodal graph over transcript, visual, topic, and entity evidence. Given a question, VideoGraph retrieves relevant nodes, expands through local graph context, and formats a compact evidence set for answer generation.

Abstract

Long videos require systems to recover sparse evidence over time from both visual and spoken content. We present VideoGraph, a training-free method that builds a persistent multimodal graph memory once per video from transcript segments, visual clips, topics, and entities. We use long video question answering as an indirect probe of whether such a memory helps recover temporally grounded evidence. VideoGraph outperforms all compared baselines, including prior graph-based methods, on NExT-QA (77.62%), EgoSchema (70.80%), and Video-MME (69.11%, medium split). The strongest supporting pattern is on NExT-QA, where gains are largest on temporal and causal questions, consistent with the intended role of the graph as a memory over temporally distributed and cross-modal evidence. Because the comparisons use different answer models and preprocessing pipelines, the results should be read as evidence that the approach is promising rather than as proof that graph memory alone explains the gains. We therefore add controlled retrieval-stage ablations and report reusable-memory footprint and tracked graph-construction API cost.

[1]Graduate Program in Computational Science, American University of Beirut, Beirut, Lebanon [2]Department of Electrical and Computer Engineering, American University of Beirut, Beirut, Lebanon. Correspondence to: Dany Chahine <ddc00@mail.aub.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Long videos make memory a central problem for multimodal systems. The evidence needed to answer a question is often sparse, distributed across time, and split between

visual events and spoken content. This same issue appears in long video evaluation and generation, where a system must preserve consistent state over entities, events, scenes, and narrative context. In this setting, entity nodes provide handles for identity drift, visual and topic nodes for scene drift, temporal and state change structure for action continuity, and transcript to visual alignment for spoken context inconsistency. In this paper, we use long video question answering as an indirect probe of memory quality: if a system can recover the right evidence at the right time, answer accuracy should improve.

Recent training-free approaches build intermediate video representations before reasoning with a large language model, including caption repositories, agentic frame selection, and tree or graph memories. However, many of these methods are query specific, mostly visual, or centered on tracked objects and frame level observations. This limits reuse across questions and leaves spoken content, topic structure, and broader multimodal context only weakly represented.

We propose **VideoGraph**, a persistent multimodal graph memory built once per video. VideoGraph parses a video into timestamped transcript segments and temporally coherent visual clips, then constructs a graph whose nodes represent transcript segments, visual clips, topics, and entities. The edges encode temporal order, transcript to visual alignment, discourse relations, and entity mentions. At inference time, a hybrid semantic and lexical retriever selects relevant seed nodes, and local graph expansion adds neighboring temporal and cross modal context before answer generation.

We evaluate VideoGraph on NExT-QA, EgoSchema, and Video-MME. The clearest gains over baselines appear on temporal and causal questions in NExT-QA, while descriptive questions show smaller relative improvements, suggesting that persistent multimodal memory helps most when the answer depends on evidence distributed across time. Because compared methods differ in backbone models and preprocessing, we report them as literature comparisons and use controlled retrieval-stage ablations to isolate the effect of graph-aware context construction.

2. Related Work

Flat retrieved evidence. Recent training-free approaches to long video question answering often build flat textual evidence before final reasoning. LLoVi (Zhang et al., 2024), LangRepo (Kahatapitiya et al., 2025), and LifelongMemory (Wang et al., 2024b) represent videos through frame or clip level language descriptions that are summarized, filtered, or retrieved for answer generation, but do not preserve explicit multimodal structure over time.

Query time adaptive evidence selection. Some methods organize evidence around the current question. VideoTree builds a query adaptive hierarchy of visual evidence (Wang et al., 2025a), while agentic systems such as VideoAgent (Wang et al., 2024a) and ViperGPT (Surís et al., 2023) iteratively select frames or compose executable reasoning steps. These systems are flexible at question time, but they do not maintain a reusable persistent memory built once per video.

Persistent memories and graph reasoning. Graph and memory based methods are closest to VideoGraph. RAVU constructs a spatio temporal entity graph from sampled frames, tracked entities, and entity level events (Malik et al., 2026). Vgent builds a reusable graph over visual clips linked by shared entities (Shen et al., 2025). GraphVideoAgent maintains an entity relation graph updated through an LLM agent loop (Chu et al., 2025). These methods show the value of explicit structure for long video reasoning, but they are primarily grounded in visual frames, clips, or tracked entities. VideoGraph instead aims to maintain a persistent multimodal graph over transcript, visual, topic, and entity evidence.

Position of this paper. VideoGraph differs from flat evidence systems by storing reusable structure, from query time adaptive systems by building the memory once per video, and from prior graph systems by incorporating spoken content and topic structure in addition to visual evidence. The contribution is architectural: integrating transcription, visual analysis, retrieval, and answer generation into a persistent multimodal graph memory rather than introducing a new model component in isolation.

3. Method

3.1. Overview

VideoGraph represents a long video as a reusable multimodal graph memory (Figure 1). Given a video V, the system first extracts temporally grounded transcript and visual evidence, then builds a graph $G = (\mathcal{N}, \mathcal{E})$ over this evidence, where $\mathcal{N}$ is the set of nodes and $\mathcal{E}$ is the set of edges. The graph stores local observations as nodes and preserves temporal, cross modal, discourse, and entity relations as edges. Given a question q, VideoGraph retrieves a compact subgraph from this memory and uses the retrieved evidence for answer generation.

3.2. Video Parsing

VideoGraph first decomposes the video into temporally coherent clips using scene based segmentation and extracts adaptive keyframes per clip (Section A.2). The audio stream is transcribed into timestamped segments. In parallel, each visual clip is analyzed using a vision language model, which

receives keyframes from the current and previous clip and outputs a clip description, salient entities, scene type, state change description, and OCR text. This produces two aligned evidence streams: transcript segments and visual clip descriptions, both grounded in video time.

3.3. Graph Memory Construction

VideoGraph converts the parsed evidence into a multimodal graph with four node types: transcript ($\mathcal{N}_T$), visual ($\mathcal{N}_V$), topic ($\mathcal{N}_P$), and entity ($\mathcal{N}_E$) nodes.

Each transcript segment becomes a transcript node with its text and timestamp. Each visual clip becomes a visual node containing the clip description, detected entities, OCR text, scene type, keyframes, and state change description. Consecutive transcript nodes are connected with temporal edges, and consecutive visual nodes are also connected with temporal edges. Transcript and visual nodes are connected with alignment edges when their time intervals overlap, allowing the graph to link what is said with what is visible at the same time.

Topic nodes are created by applying an LLM discourse analysis step to the transcript nodes. The model groups related segments into topics, returning a title, description, keywords, and member nodes. A topic node covers the time range of its members. The same step identifies inter-segment relations (explanation, causality, temporal order, contrast), added as typed edges.

Entity nodes are created by extracting canonical entities (name, type, aliases) from the combined transcript and visual evidence using an LLM. Each entity is grounded by matching against transcript and visual nodes, and linked to nodes in which it is mentioned or depicted, connecting evidence about the same entity across modalities and time.

3.4. Graph Retrieval

For answering, VideoGraph retrieves evidence from the graph rather than passing the full video representation to the language model. Each node is converted into a textual representation (combining its relevant fields) and embedded.

Given a question q and graph node n, VideoGraph computes

$$s(q, n) = \alpha\, s_{\text{emb}}(q, n) + (1 - \alpha)\, s_{\text{lex}}(q, n),$$

where s_{emb} is clamped cosine similarity and s_{lex} is normalized word overlap. Transcript nodes receive a small relevance boost controlled by $\beta > 1$ (full formulas in Section A.1).

The highest scoring nodes form the initial evidence. VideoGraph then expands through local graph neighborhoods, following temporal and alignment edges to include neighboring moments and cross modal context.

State change information serves as a separate retrieval channel for transition questions. The retrieved subgraph is formatted as structured textual evidence, with primary nodes separated from expanded context, and passed to the answer generator.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate on three VideoQA benchmarks that are often included in long-video papers: EgoSchema (Mangalam et al., 2023), NExT-QA (Xiao et al., 2021), and Video-MME (Fu et al., 2025). We report results on the EgoSchema subset, NExT-QA validation split, and the short, medium, and long splits of Video-MME. Together, these cover temporal and causal reasoning, egocentric long horizon understanding, and robustness across video duration.

Implementation Details. We use GPT-4o for visual analysis and answer generation, Whisper-1 for transcription, and `text-embedding-3-small` for graph retrieval. Videos are segmented with PySceneDetect, followed by adaptive keyframe extraction and OCR. Additional implementation settings, retrieval hyperparameters, scene sampling details, and the answer prompt are reported in Section A.2 (for Video-MME, we use Whisper-transcribed audio rather than the benchmark-provided subtitles). Code is available at `https://github.com/VideoGraph-AUB/videograph`.

For NExT-QA, we compare against caption and retrieval based methods (Yu et al., 2023; Surís et al., 2023; Zhang et al., 2024), agent based methods (Wang et al., 2024a), graph based methods (Chu et al., 2025; Malik et al., 2025), and hierarchical video representations (Wang et al., 2025a). Baseline sourcing details are reported in Section A.2. Since compared systems differ in answer models, preprocessing, and auxiliary components, the tables below provide literature context rather than controlled comparisons.

Table 1. NExT-QA validation accuracy in the literature comparison. The question type breakdown is reported in Table 5.

Method	(M)LLM	Acc
LLoVi (Zhang et al., 2024)	GPT-4	73.8
SeViLA (Yu et al., 2023)	Flan-T5	63.6
ViperGPT (Surís et al., 2023)	GPT-3	60.0
VideoAgent (Wang et al., 2024a)	GPT-4	71.3
GraphVideoAgent (Chu et al., 2025)	GPT-4	73.3
RAVU (Malik et al., 2025)	Gemini 1.5 Flash	71.93
VideoTree (Wang et al., 2025a)	GPT-4	75.6
VideoGraph (ours)	GPT-4o	**77.62**

Table 2. EgoSchema subset accuracy in the literature comparison.

Method	(M)LLM	Acc
LLoVi (Zhang et al., 2024)	GPT-4	61.2
SeViLA (Yu et al., 2023)	Flan-T5	25.7
VideoAgent (Wang et al., 2024a)	GPT-4	60.2
GraphVideoAgent (Chu et al., 2025)	GPT-4	62.7
RAVU (Malik et al., 2025)	Gemini 1.5 Flash	66.6
VideoTree (Wang et al., 2025a)	GPT-4	66.2
LifelongMemory (Wang et al., 2024b)	GPT-4	68.0
LifelongMemory (Wang et al., 2024b)	GPT-4o	70.6
VideoGraph (ours)	GPT-4o	**70.80**

Table 3. Video-MME medium split accuracy in the literature comparison. The full duration breakdown is reported in Table 6.

Method	(M)LLM	Acc
LLoVi† (Zhang et al., 2024)	GPT-4	53.2
VideoTree† (Wang et al., 2025a)	GPT-4	59.9
VideoGraph (ours)	GPT-4o	**69.11**

†As reported in (Wang et al., 2025b).

5. Results and Analysis

Overall comparison. VideoGraph outperforms all compared baselines, including prior graph-based methods, on each selected split: 77.62 on NExT-QA validation, 70.80 on the EgoSchema subset, and 69.11 on the Video-MME medium split.

Where gains appear. The clearest pattern appears on NExT-QA, using the question type breakdown in Table 5. VideoGraph reaches the strongest temporal accuracy (72.98) and causal accuracy (79.40), while VideoTree remains strongest on descriptive questions (83.9 vs. 81.98). This suggests that the largest relative gains are concentrated on questions that require evidence across time or across causally linked moments, rather than on purely descriptive visual questions. The same trend is consistent with the broader benchmark results: VideoGraph is comparable to the strongest single-run EgoSchema baseline, LifelongMemory with GPT-4o (70.6 vs. 70.80), and improves over VideoTree on the Video-MME medium split by 9.21 points, with additional short and long split results reported in Table 6.

Table 4. NExT-QA retrieval-stage ablations. All variants use the same precomputed graph artifacts, answer model, and prompt, and change only the retrieved context.

Variant	AccT	AccC	AccD	Acc
Full VideoGraph	72.98	79.40	81.98	77.62
No graph expansion	69.58	77.75	81.60	75.56
Flat retrieval	69.19	77.10	83.01	75.36
Transcript only	48.91	59.84	45.43	54.08
Visual only	71.37	78.29	82.50	76.56

Interpreting the design. Table 4 isolates retrieval and context construction after the graph has been built. Removing graph expansion drops overall accuracy by 2.06 points, while flattening the memory into top-k node retrieval drops accuracy by 2.26 points. Both drops are largest on temporal questions, matching the role of graph expansion in recovering nearby and cross-modal context. Transcript-only context performs much worse, showing that visual evidence is essential. Visual-only context remains strong and is slightly better on descriptive questions, which suggests that localized descriptive questions may benefit from more visual-focused or question-type-adaptive retrieval. On Video-MME medium, the same trend holds: no-expansion and flat retrieval reduce accuracy from 69.11 to 63.67 and 61.33, respectively (Table 7).

Persistent memory and reuse. VideoGraph constructs the graph once per video and reuses it across questions. From the completed evaluation artifacts, the same graph is reused across 8.76 questions/video on NExT-QA and 3 questions/video on each Video-MME split; EgoSchema has one question/video and therefore measures footprint rather than query amortization. The reusable QA memory, counted as the graph, node embeddings, visual-channel embeddings, and state-change descriptions, has median storage of 1.05 MiB/video on NExT-QA, 3.29 MiB/video on EgoSchema, and 2.04/12.02/25.35 MiB/video on Video-MME short/medium/long. On Video-MME, median graph size grows from 42 nodes and 128 edges on short videos to 801 nodes and 2560 edges on long videos, while median reusable memory on long videos remains 25.35 MiB/video compared with 206.29 MiB/video for the raw videos. In a tracked five-video sample from each Video-MME split with cache disabled, median graph-construction API cost was $0.092/video, $0.447/video, and $2.284/video for short, medium, and long videos, respectively; this cost is paid once per video before reuse.

6. Discussion and Limitations

Strengths. VideoGraph is competitive in literature comparisons, with the clearest gains on temporal and causal questions in NExT-QA, consistent with the intended role of persistent multimodal memory.

Limitations. The external benchmark tables remain heterogeneous literature comparisons: methods use different answer models, captioners, prompts, and preprocessing pipelines. Our ablations keep the VideoGraph pipeline fixed, but they do not provide a full controlled head-to-head comparison against reimplemented baselines. The datasets used here do not provide per-question supporting clips or timestamps, so we cannot directly report retrieval recall; question answering accuracy remains an indirect probe of

memory quality. Finally, the descriptive-question results and the visual-only ablation suggest a path for future work: question-type-adaptive retrieval that gives more weight to localized visual/OCR evidence while preserving the process-once, reuse-many graph design.

Outlook. These results support the utility of persistent multimodal memory for long video question answering, particularly for temporal and causal reasoning. Isolating the contribution of the graph structure from other system differences remains an open question. Beyond QA, the same representation is relevant to long-horizon video generation and evaluation: transcript, visual, topic, entity, temporal, and state-change nodes provide an inspectable record of what happened, when it happened, and which entities or scenes were involved. Such a memory could support audits of generated videos for identity drift, action continuity, scene consistency, and speech-visual mismatch. We view QA accuracy as one measurable proxy for this broader problem of reliable long-horizon video memory, rather than as the only downstream use of the graph.

7. Conclusion

VideoGraph addresses long video question answering with a persistent multimodal graph memory built from transcript, visual, topic, and entity evidence. It outperforms all compared baselines, including prior graph-based methods, on NExT-QA, EgoSchema, and Video-MME. The strongest supported pattern is narrower: the largest relative gains appear on temporal and causal questions in NExT-QA, while descriptive questions show smaller improvements over baselines.

The results support the view that persistent multimodal memory can help long video question answering, especially when the answer depends on temporally distributed or cross-modal evidence. The added ablations support the value of graph-aware context over flat retrieval, while direct evidence annotations and fully controlled baseline reimplementations remain important future work.

Impact Statement

This work studies structured memory for long horizon video understanding. By organizing transcripts, visual descriptions, OCR, and entity information from long videos, such systems can make evidence easier to retrieve and inspect. This may improve transparency in video question answering, but could also make sensitive information easier to search or aggregate when applied to personal or private footage. We therefore emphasize careful dataset selection, privacy aware use, and transparent reporting of retrieved evidence and system limitations.

References

Chu, M., Li, Y., and Chua, T.-S. Understanding long videos via LLM-powered entity relation graphs, 2025.

Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., and Sun, X. 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, June 2025.

Kahatapitiya, K., Ranasinghe, K., Park, J., and Ryoo, M. S. Language repository for long video understanding. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5627–5646, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.294.

Malik, S., Yamada, M., Singh, A., and Aggarwal, D. RAVU: Retrieval augmented video understanding with compositional reasoning over graph, 2025. URL https://arxiv.org/abs/2505.03173.

Malik, S., Singh, A., Yamada, M., and Aggarwal, D. RAVU: Retrieval augmented video understanding with compositional reasoning over graph. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2869–2878, March 2026.

Mangalam, K., Akshulakov, R., and Malik, J. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244. Curran Associates, Inc., 2023.

Shen, X., Zhang, W., Chen, J., and Elhoseiny, M. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. In *Advances in Neural Information Processing Systems*, 2025.

Surís, D., Menon, S., and Vondrick, C. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, October 2023. doi: 10.1109/ICCV51070.2023.01092.

Wang, X., Zhang, Y., Zohar, O., and Yeung-Levy, S. VideoAgent: Long-form video understanding with large language model as agent. In *Computer Vision – ECCV 2024*, pp. 58–76. Springer Nature Switzerland, 2024a. doi: 10.1007/978-3-031-72989-8_4.

Wang, Y., Yang, Y., and Ren, M. LifelongMemory: Leveraging LLMs for answering queries in long-form egocentric videos, 2024b.

Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., and Bansal, M. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3283, June 2025a. doi: 10.1109/CVPR52734.2025.00311.

Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., and Bansal, M. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos, 2025b. URL https://arxiv.org/abs/2405.19209.

Xiao, J., Shang, X., Yao, A., and Chua, T.-S. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, June 2021.

Yu, S., Cho, J., Yadav, P., and Bansal, M. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, 2023.

Zhang, C., Lu, T., Islam, M. M., Wang, Z., Yu, S., Bansal, M., and Bertasius, G. A simple LLM framework for long-range video question-answering. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1209.

A. Supplementary Material

A.1. Retrieval Scoring

We provide the full retrieval scoring formulas referenced in Section 3.4. Given a question q and graph node n, VideoGraph computes

$$s(q, n) = \alpha\, s_{\text{emb}}(q, n) + (1 - \alpha)\, s_{\text{lex}}(q, n),$$

where α controls the contribution of semantic similarity relative to lexical overlap. The embedding score is cosine similarity clamped to the range $[0, 1]$:

$$s_{\text{emb}}(q, n) = \max\left(0, \cos(\mathbf{e}_q, \mathbf{e}_n)\right),$$

where $\mathbf{e}_q$ and $\mathbf{e}_n$ are the embedding vectors of the question and node text, respectively. The lexical score is normalized word overlap:

$$s_{\text{lex}}(q, n) = \frac{|W(q) \cap W(n)|}{|W(q)|},$$

where $W(q)$ and $W(n)$ are the sets of words in the question and node text. For transcript nodes, VideoGraph applies a small relevance boost:

$$s(q, n) \leftarrow \min(1, \beta \cdot s(q, n)),$$

where $\beta > 1$ controls the boost strength. In our experiments, we use $\alpha = 0.7$ and $\beta = 1.1$.

A.2. Implementation Settings and Prompts

Baseline sourcing. Unless otherwise noted, baseline results are taken from the main results in the corresponding papers. For LLoVi (Zhang et al., 2024) on NExT-QA, we report the CogAgent captioner setting from the main LLoVi paper. For RAVU (Malik et al., 2025), we report the overall accuracies from the authors' arXiv version, where Gemini blocked samples are counted as incorrect; the WACV paper reports results only on the non-blocked subset (Malik et al., 2026). For Video-MME, VideoTree's (Wang et al., 2025a) long split result is reported in the main CVPR paper, as is the LLoVi long split result reported there for comparison. The VideoTree short and medium split results and the LLoVi short and medium split results are taken from the VideoTree arXiv supplementary material (Wang et al., 2025b). For LifelongMemory (Wang et al., 2024b) on EgoSchema, we report the single-run result (70.6 with GPT-4o); the authors also report a vote-by-confidence ensemble variant (72.0), which we exclude for consistency with single-run comparisons. Since prior work uses different preprocessing, prompting, and auxiliary models, we report GPT-4 or GPT-4 class configurations when available.

Models and decoding. VideoGraph uses GPT-4o for question answering, visual captioning, and OCR, Whisper-1 for transcription, and `text-embedding-3-small` for node embeddings. The question-answering temperature and visual-captioning temperature are both set to 0.0. Topic and entity extraction during graph construction uses temperature 0.3.

Retrieval settings. The retrieval score uses $\alpha = 0.7$ for the semantic–lexical interpolation and $\beta = 1.1$ for the transcript relevance boost. The initial retrieved set contains the top 7 seed nodes, and graph expansion uses one hop from the seed nodes.

Video segmentation and keyframes. Videos are segmented using PySceneDetect's AdaptiveDetector with adaptive threshold 0.5, minimum scene length 2.5 seconds, and window width 2. For each scene clip, internal representative keyframes are sampled adaptively: clips of at most 8 seconds use one internal keyframe, clips of at most 20 seconds use two internal keyframes, and clips longer than 20 seconds use three internal keyframes. For n internal keyframes, frame i is sampled at position

$$\frac{i+1}{n+1} \cdot \text{clip duration}.$$

The pipeline also adds one boundary keyframe near the end of each clip, so the usual total is two frames for short clips, three for medium clips, and four for long clips.

Answer prompt. For multiple-choice answering, VideoGraph uses the following system prompt:

```
You are answering a multiple-choice question about a video.
Based on the provided context from the video's
knowledge graph, select the best answer.
Respond with ONLY the option number ({valid_indices}). Nothing else.
```

The user prompt template is:

```
Context from video:
{retrieved_context}

Question: {question}

Options:
0: {option_0}
1: {option_1}
...

Answer (0-{N-1}):
```

For EgoSchema, the prompt additionally inserts the note:

```
Note: In this video, 'C' refers to the camera wearer (first-person view) and
'O' refers to another person.
```

A.3. Detailed Results

Table 5. NExT-QA validation accuracy by question type. AccT, AccC, and AccD denote temporal, causal, and descriptive accuracy.

Method	(M)LLM	AccT	AccC	AccD	Acc
LLoVi (Zhang et al., 2024)	GPT-4	70.2	73.7	81.9	73.8
SeViLA (Yu et al., 2023)	Flan-T5	61.3	61.5	75.6	63.6
ViperGPT (Surís et al., 2023)	GPT-3	–	–	–	60.0
VideoAgent (Wang et al., 2024a)	GPT-4	64.5	72.7	81.1	71.3
GraphVideoAgent (Chu et al., 2025)	GPT-4	65.2	74.6	83.5	73.3
RAVU (Malik et al., 2025)	Gemini 1.5 Flash	66.56	74.4	74.64	71.93
VideoTree (Wang et al., 2025a)	GPT-4	70.6	76.5	**83.9**	75.6
VideoGraph (ours)	GPT-4o	**72.98**	**79.40**	81.98	**77.62**

Table 6. Video-MME accuracy by duration split.

Method	(M)LLM	Short	Medium	Long
LLoVi† (Zhang et al., 2024)	GPT-4	62.1	53.2	48.8
VideoTree† (Wang et al., 2025a)	GPT-4	67.8	59.9	54.2
VideoGraph (ours)	GPT-4o	**75.78**	**69.11**	**60.08**

†Short and medium splits as reported in supplementary material of (Wang et al., 2025b); long split from (Wang et al., 2025a).

Table 7. Video-MME medium retrieval-stage ablations. All variants use the same evaluation split and precomputed graph artifacts.

Variant	Acc	Delta
Full VideoGraph	69.11	–
No graph expansion	63.67	-5.44
Flat retrieval	61.33	-7.78
Transcript only	54.78	-14.33
Visual only	64.22	-4.89
Temporal-only expansion	65.67	-3.44
Alignment-only expansion	68.67	-0.44