Graph Against the Machine: a Neuro-Symbolic Approach for Enhanced Video Question Answering

Fabio Lusha[a†], **Agnese Chiatti**[a‡], **Sara Pidò**[a‡], **Nico Catalano**[a‡] **and Matteo Matteucci**[a‡]

[a]Politecnico di Milano, Italy. †{name.surname}@mail.polimi.it, ‡{name.surname}@polimi.it

Abstract.

Video Question Answering (VideoQA) is a key problem contributing to advanced video understanding. The rise of Multimodal Large Language Models (MLLMs) has accelerated the improvement on VideoQA tasks. However, MLLMs can produce inconsistent output even for similar prompts and suffer from hallucinations and biases. In this position paper, we envisage a novel pipeline, where scene graphs representing people, objects, and relationships in a video are injected in the MLLM prompt. We hypothesise that leveraging a symbolic representation of the video content can improve accuracy and verifiability and reduce the latency of MLLMs for VideoQA.

1 Introduction and Motivation

The rapid growth in the production and indexing of video content across virtually all industry sectors - ranging from healthcare and law enforcement to education and entertainment - calls for effective and trustworthy methods for autonomously understanding videos. Automating video understanding could support application scenarios of significant social impact, such as accident detection and diagnosis in autonomous driving [6, 5], or fall detection and behaviour monitoring for fragile and elderly patients [14, 15].

However, video understanding is a challenging task for state-of-the-art methods in Computer Vision as it requires advanced spatiotemporal, causal, and abductive reasoning capabilities. Video Question Answering (VideoQA), i.e., the ability to autonomously answer natural language queries about an input video, is one crucial prerequisite towards achieving advanced video understanding [19, 4]. This problem is particularly challenging in the case of long-form video clips [4], where models ought to go beyond frame-level comprehension to grasp long-range dependencies and complex interactions between people and objects.

The rise of Multimodal Large Language Models (MLLMs) has expedited the advancement on VideoQA thanks to the impressive accuracy of these models in answering queries from multi-modal prompts comprising video and text [17]. However, this rapid advancement has also raised significant concerns. First, these models operate as black-boxes and produce inconsistent outputs for similar prompts, complicating the task of verifying answers against supporting evidence [13]. Second, they often hallucinate, fabricating objects, people, and events inconsistent with the video content [1]. Moreover, they frequently over-rely on the textual prompt, neglecting the visual input - an issue also known, in the literature, as language bias [10, 8]. This issue is exacerbated by the MLLMs potential to produce harmful, discriminatory, or toxic content [8].

Scene Graph Generation (SGG), which extracts spatio-temporal graphs from videos to represent entities and their relationships, has been increasingly overshadowed by the rise of MLLMs. However, scene graphs can help structure and make more consistent the MLLMs responses, while offering a graphical aid to explain model answers. For example spatio-temporal scene graphs can provide timestamped links between “pedestrian,” “vehicle,” and “crosswalk”, allowing instantaneous path-finding analyses of collision sequences. Similarly, in assisted living scenarios, a graph could be used to investigate the cause of a fall accident, by analysing that, e.g., “a liquid substance" appeared “on the floor" just before the “fall" event. Crucially, because the heavy lifting of video parsing is done just once, every subsequent query runs directly against the graph, which provides a lightweight representation of the video.

Focus and background We propose to adopt scene graph representations as a bridge between the visual content of the video and the textual query. This hybrid approach is aligned with the rapidly re-emerging interest in the field of Neuro-symbolic (NeSy) AI, which advocates for leveraging the strengths of sub-symbolic (i.e., data-driven) learning methods and symbolic knowledge representations [16, 9]. Despite continuous efforts in the fields of SGG and NeSy AI, the integration of scene graphs in prompts remains rather unexplored in the Computer Vision community. A few recent approaches to VideoQA based on LLMs exploit scene graphs i) only for specific sub-tasks such as object tracking or action recognition, ii) by adopting expensive training procedures to fine-tune the model directly on graph data [7, 18, 2]. We explore instead a different approach where scene graphs are injected directly in prompts, inspired by promising results tested in [12] on images, showing that integrating scene graphs in textual prompts can improve the compositional reasoning abilities of MLLMs. Our setting is similar, in principle, to Chain of Thought reasoning [3, 20, 7], where a more complex problem is broken down into individual subproblems. That is, we aim at encouraging the MLLM to think about the graph structure before providing answers. To achieve this objective, we ask:

- *Can MLLMs be effectively applied to generate scene graphs from video inputs without resorting to manual annotations?*
- *Can integrating scene graphs into textual prompts in place of video frames improve the MLLMs accuracy and inference speed on VideoQA tasks?*

2 Proposed approach

In our experiments we compare two pathways: generative graph QA and direct VideoQA (Figure 1). Both pipelines share common upstream components, but diverge in how they represent and process visual information.

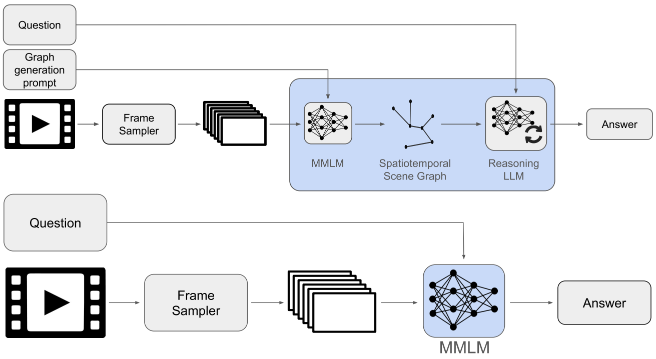

Figure 1. Overview of our two VideoQA pipelines. **Top:** Generative Graph QA: frames are first converted to scene graphs by an MLLM, aggregated, and then reasoned over by an LLM. **Bottom:** Direct VideoQA: frames are fed directly to an MLLM for VideoQA.

Temporal Sampling Given a video of duration T, we applied two different sampling strategies: i) using a fixed number of frames (N=5) at uniform intervals, and ii) uniform sampling at 1 fps.

Prompt Engineering for Structured Perception and Reasoning To ensure consistent behavior across models, we adopt a two-stage prompting scheme based on established zero-shot Chain-of-Thought principles [11]. Each stage of our pipeline is driven by:

Task-Prompting. For Generative Graph QA, we use a prompt comprising the graph, question, and answer alternatives. For Direct VideoQA, in the prompt we replace the graph with the image.

Output Formatting. We constrain responses to structured graphs or explicit final answers, to aid deterministic parsing and make the overall evaluation more consistent.

NeSy Path: Generative Graph QA In the NeSy pipeline, we decouple perception and reasoning through an explicit scene-graph intermediate representation. This process unfolds in three stages:

Scene-Graph Generation is performed in two variants: per-frame generation, where each frame is independently processed into a scene graph; batch generation, where all frames are jointly presented to the MLLM, to provide spatial and temporal context.

Graph Aggregation. We concatenate the per-frame graphs into a single graph G=(V,E), i.e., a temporally-ordered list.

Graph-Based QA. The LLM is fed with the graph, the question, and multiple-choice alternatives. The LLM reasons step-by-step over the structured input to select the most consistent answer. The presence of the graph enforces transparency over the reasoning process, supporting error analysis and human verification.

Baseline Path: Direct VideoQA In contrast, in the Direct VideoQA pipeline frames are embedded directly into a multimodal prompt alongside the question and answer choices. The MLLM is responsible for both perception and reasoning, leveraging its joint representation space to produce an answer. Operating only in one step, this approach lacks transparency and modularity, and may underperform especially on clips of growing length and complexity.

Evaluation Protocols To rigorously compare these paradigms, we implement both pipelines under identical sampling strategies and prompt templates, using Gemma3 4b as both the MLLM and LLM. All components are run on a single NVIDIA GTX 1080 GPU. We evaluate on 1,048 questions in the validation set of STAR [19], with respect to the following metrics: (i) Accuracy: Exact-match correctness; (ii) Latency: End-to-end inference time per question; (iii) Complementarity: Unique hits, instances where one pipeline succeeds and the other fails. Table 1 shows our preliminary results.

Pipeline	Int.	Seq.	Pre.	Fea.	Avg	Lat.(h)	U.H.(%)
Direct VideoQA	48.0	51.5	41.8	39.1	**45.7**	14	–
GT GraphQA	58.0	63.3	75.6	66.1	**62.8**	3	–
Gen-Graph (Frame)	40.3	35.4	39.3	32.7	**37.1**	48+5*	12.3
Gen-Graph (Batch)	31.0	26.1	27.6	26.4	**27.9**	16+4*	8.6

Table 1. Accuracy per question type. Latency in minutes. U.H.: Unique Hits—examples correctly answered only by the Gen-Graph pipeline. STAR data includes four subsets: Interaction (Int.), Sequence (Seq.), Prediction (Pre.), Feasibility (Fea.). *SGG latency + Graph-Based QA latency

3 Discussion and Future Directions

Our dual-pipeline study highlights a central tradeoff in the design of VideoQA systems: efficiency and interpretability via symbolic representations versus end-to-end accuracy from direct multimodal inference. As shown in Table 1, methods based on generated graphs exhibit lower overall accuracy compared to the direct VideoQA approach. Crucially, using ground truth scene graphs provided with STAR (GT GraphQA, in the table) led to the highest accuracy for all question types. Thus, we can hypothesise that the 8–10 % drop in GT GraphQA experiments is primarily caused by the generated graphs rather than the QA step. For all graph-based pipelines, the latency is higher than in the Direct VideoQA setting, due to the computational cost of scene graph generation (also note the lowest latency in the GT GraphQA case). Despite this overhead, the symbolic pipeline recovers 12% of questions that the direct pipeline fails to answer correctly. Moreover, thanks to integrating graphs in prompts, the NeSy pipelines produce outputs ready for human verification, enabling the direct comparison between answers and graphs.

We also observe meaningful performance differences between NeSy variants. Batch-based graph extraction, which incorporates temporal context across frames, yields slightly lower accuracy but lower latency thanks to generating graphs once per batch. Hence, how symbolic information is extracted and structured directly influences downstream performance and latency.

In response to our initial research questions: Neuro-symbolic integration of graphs in MLLM prompts can enhance accuracy and latency in sub-symbolic pipelines when relying on ground truth graphs. However, a significant performance gap remains when graphs are generated with LLMs. These findings suggest that scene graphs can serve as interpretable interfaces within modern MLLMs pipelines.

Limitations and Future Work Our current implementation is still preliminary and suffers from a few limitations. First, generated graphs omit attributes or objects essential for answering certain questions. Second, generating graphs and answering questions in two steps causes a bottleneck. Finally, we evaluated only on short STAR clips and generalization to longer videos is still untested.

However, we see several promising directions for future exploration:

- Improving the quality of generated graphs by dynamically sampling the most salient video frames, as well as ensuring temporal alignment and co-reference resolution.
- Extending our prompting strategies by combining graphs and frames in the same prompt, enabling LLMs to cross-reference symbolic and raw visual data.
- End-to-End Learning, explore graph-aware finetuning of VLMs for joint SGG and VideoQA.
- Extending our evaluation to transparency and trustworthiness metrics, and conducting statistical robustness tests to verify complementarity effects across datasets and question types.

Acknowledgements

This work has been supported by Politecnico di Milano through the 2024 MSCA Seal of Excellence fellowship (project ReFiNe) and by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence).

References

[1] K. Bae, J. Kim, S. Lee, S. Lee, G. Lee, and J. Choi. MASH-VLM: Mitigating action-scene hallucination in Video-LLMs through disentangled spatial-temporal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 13744–13753, 2025.

[2] Z. Bai, R. Wang, D. Gao, and X. Chen. Event graph guided compositional spatial–temporal reasoning for video question answering. *IEEE Transactions on Image Processing*, 33:1109–1121, 2024.

[3] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690, 2024.

[4] J.-J. Chen, Y.-C. Liao, H.-C. Lin, Y.-C. Yu, Y.-C. Chen, and Y.-C. F. Wang. ReXTime: A Benchmark Suite for Reasoning-Across-Time in Videos. *arXiv:2406.19392*, 2024.

[5] J. Fang, J. Qiao, J. Xue, and Z. Li. Vision-based traffic accident detection and anticipation: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):1983–1999, 2023.

[6] J. Fang, L.-l. Li, J. Zhou, J. Xiao, H. Yu, C. Lv, J. Xue, and T.-S. Chua. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22030–22040, June 2024.

[7] H. Fei, S. Wu, W. Ji, H. Zhang, M. Zhang, M. L. Lee, and W. Hsu. Video-of-thought: step-by-step video reasoning from perception to cognition. In *Proceedings of the 41st International Conference on Machine Learning*, pages 13109–13125, 2024.

[8] Y. Gou, K. Chen, et al. Eyes closed, safety on: Protecting Multimodal LLMs via image-to-text transformation. In *In Proceedings of the European Conference of Computer Vision (ECCV)*, 2024.

[9] P. Hitzler, A. Dalal, M. S. Mahdavinejad, and S. S. Norouzi, editors. *Handbook on Neurosymbolic AI and Knowledge Graphs*. Frontiers in Artificial Intelligence and Applications. IOS Press, 2025.

[10] T. Huai, S. Yang, J. Zhang, J. Zhao, and L. He. Debiased Visual Question Answering via the perspective of question types. *Pattern Recognition Letters*, 178:181–187, 2024. ISSN 0167-8655. doi: 10.1016/j.patrec.2024.01.009.

[11] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

[12] C. Mitra, B. Huang, T. Darrell, and R. Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14420–14431, 2024.

[13] H. Qiu, M. Gao, L. Qian, K. Pan, Q. Yu, J. Li, W. Wang, S. Tang, Y. Zhuang, and T.-S. Chua. STEP: Enhancing Video-LLMs' compositional reasoning by spatio-temporal graph-guided self-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3284–3294, 2025.

[14] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris, and G. Farias. Fall detection and activity recognition using human skeleton features. *Ieee Access*, 9:33532–33542, 2021.

[15] L. Romeo, R. Marani, T. D'Orazio, and G. Cicirelli. Video based mobility monitoring of elderly people using deep learning models. *IEEE Access*, 11:2804–2819, 2023.

[16] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler. Neuro-symbolic artificial intelligence: Current trends. *AI Communications*, 34(3):197–209, 2022.

[17] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu, et al. Video understanding with Large Language Models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[18] A. Urooj, H. Kuehne, B. Wu, K. Chheu, W. Bousselham, C. Gan, N. Lobo, and M. Shah. Learning situation hyper-graphs for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14879–14889, June 2023.

[19] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *In Proceedings of NeurIPS*, 2021.

[20] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.