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

Fabio Lusha[1,*], Agnese Chiatti[1], Sara Pidò[1], Nico Catalano[1] and Matteo Matteucci[1]

[1]Artificial Intelligence and Robotics Lab (AIRLab), Politecnico di Milano, Italy.

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 [1, 2], or fall detection and behaviour monitoring for fragile and elderly patients [3, 4].

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 [5, 6]. This problem is particularly challenging in the case of long-form video clips [6], 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 [7]. 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 [8]. Second, they often hallucinate, fabricating objects, people, and events inconsistent with the video content [9]. Moreover, they frequently over-rely on the textual prompt, neglecting the visual input - an issue also known, in the literature, as language bias [10, 11]. This issue is exacerbated by the MLLMs potential to produce harmful, discriminatory, or toxic content [11].

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 [12, 13]. 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 [14, 15, 16]. We explore instead a different approach where scene graphs are injected directly in prompts, inspired by promising results tested in [17] 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 [18, 19, 14], 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?*

ANSyA 2025: 1st International Workshop on Advanced Neuro-Symbolic Applications, co-located with ECAI 2025.

*Corresponding author.

fabio.lusha@mail.polimi.it (F. Lusha); agnese.chiatti@polimi.it (A. Chiatti); sara.pido@polimi.it (S. Pidò); nico.catalano@polimi.it (N. Catalano); matteo.matteucci@polimi.it (M. Matteucci)

https://achiatti.github.io/ (A. Chiatti); https://nicocatalano.github.io/ (N. Catalano)

0000-0003-3594-731X (A. Chiatti); 0000-0003-1425-1719 (S. Pidò); 0009-0004-2731-1068 (N. Catalano); 0000-0002-8724-2859 (M. Matteucci)

1.1. Neuro-symbolic AI

The interchange of symbolic and sub-symbolic approaches is not new: in fact, it has characterised AI development since the earliest days of the field [20]. Historically, there has been a tension between these two approaches, for instance with the resurgence of Neural Networks in the 2010s, taking over "first-wave" systems based on logic programming and Bayesian inference. Nevertheless, there have also been synthesis phases, like the proposal of neuro-fuzzy systems in the 1990s [21]. AI experts increasingly agree that adopting a Neuro-symbolic (NeSy) approach can overhaul this synthesis, achieving the best of both worlds [20, 13, 12]. Different NeSy methods have been most recently proposed that are specifically tailored to Neural Networks. In this context, symbolic knowledge can be integrated at different levels [22]. (i) In pre-processing, to augment the training examples or to partition the learning space. (ii) Within the intermediate layers of the Network, or (iii) as a part of the architectural topology, or (iv) in the optimisation function. (v) In the post-processing stages, to validate the model predictions. Methods in groups (i), (iii), (v) use knowledge to bias the learning through structural constraints, providing an increased control on the reasoning process compared to models in (ii) and (iv), where symbolic knowledge is only approximately satisfied as part of the learning objectives, trading off interpretability with inference scalability [23]. In our approach, we propose to inject structured scene graphs in Video-LLM prompts (i.e., in pre-processing) before asking the model to solve VideoQA problems.

1.2. Scene Graph Generation

Scene Graph Generation (SGG) broadly refers to the process of identifying key entities within an image or video frame and representing them as a structured graph, where objects are nodes and their relationships are depicted as edges. Nodes and edges can be further enriched with attributes - e.g., the objects size and colour. To keep track of the temporal evolution of elements and events in videos, an extension of this representation has been proposed known as Spatio Temporal Scene Graph (STSG). The most common approach to constructing STSGs is maintaining one scene graph per video frame and connecting successive graphs as a temporally ordered sequence [24, 5]. In this context, hypergraph can be used to encode higher-order visual relationships across frames, as in the case of the Situated Reasoning in Real-World Videos (STAR) [5] and Video Scene Graph Reasoning (VSGR) [25] datasets. With the emergence of Video-LLM methods showing impressive zero-shot learning capabilities even on unseen tasks (Video-LLaMA [26], Video-LLaVA [27], Gemma [28], Video-ChatGPT [29]) an increasing number of works resorts to Video-LLMs to autonomously construct scene graphs from video frames [25, 8]. While this approach overcomes the cost of manually curating rich graph representations, it can introduce errors and bias in the SGG process. In this work, we explore the use of Video-LLMs for Scene Graph Generation on the STAR dataset, which conveniently provides ground truth scene graphs and is targeted at solving advanced VideoQA tasks that require situational reasoning capabilities.

1.3. Video Question Answering

Video Question Answering (VideoQA) asks a model to perceive and interpret the content of a video and then answer natural-language queries about it. Questions can range from simple descriptions of what appears in individual frames to reasoning about temporal aspects, causality or intentionality based on commonsense and prior knowledge. In the *closed-form* setting the model is given a question plus a small set of answer alternatives and must select the correct one. This setup simplifies the evaluation and the application of a randomised baseline method for comparison. However, it also allows the model to exploit text bias and guess the correct answer from wording patterns in the provided answers rather than truly understanding the video. In the *open-ended* setting the model must generate a free-form response with no predefined choices. This configuration ensures to assess more rigorously the video understanding capabilities of a model. However, it requires flexible evaluation metrics to account for semantic and syntactic variability in model answers. Benchmark datasets reflect these variants and levels of difficulty: VLEP [30], STAR [5] and IntentQA [31] present multiple-choice event prediction, stepwise reasoning and intent inference tasks; Social-IQ [32], Causal-VidQA [33] and NExT-QA [34] offer open-ended questions that probe social understanding, causal and counterfactual inference and fine-grained temporal action reasoning.

Video question answering has been a longstanding challenge in Computer Vision and AI, with early approaches using cross-modal attention [35, 36, 37] motion-appearance memory [38, 39, 40], and Graph Neural Networks [41, 42, 43] to model interactions in video sequences. These methods often struggled with long videos where multiple objects and actions interact over time, leading to confusion between similar relations or failure to capture dynamic events. The recent advent of Multimodal Large Language Models (MLLMs) has revitalized research in this area by providing powerful pretrained backbones capable of reasoning over both visual and textual inputs. Some pioneering MLLM-based methods, such as TOPA [44], VideoChat2 [45] and the Look-Remember-Reason framework [46], extend language models to video understanding using text-only pre-alignment strategies or low-level surrogate tasks to ground their predictions. In parallel, a new wave of research is integrating explicit graph representations into MLLMs to further enhance spatiotemporal reasoning. MotionEpic [14] incorporates a spatial-temporal scene graph for pixel-level grounding, the SHG-VideoQA model [15] predicts situation hypergraphs to capture actions and object relationships, and the HSST approach [16] builds hierarchical event graphs spanning objects, relations, scenes and actions, together with a spatial-temporal transformer that exploits edge-guided attention for compositional reasoning.

1.4. Chain-of-thought Prompting

Different representations have been proposed that organise LLM responses, i.e., "thoughts", so as to guide the decomposition of complex reasoning problems into smaller subtasks, an approach also known, in the literature, as Chain-of-Thought (CoT) reasoning. In the Tree-of-Thoughts (ToT) approach [19], the model responses are displayed as a tree, to keep track of multiple thought chains. The Graph of Thoughts framework (GoT) [18] further extends ToT by organising prompts and responses as directed acyclic graphs.

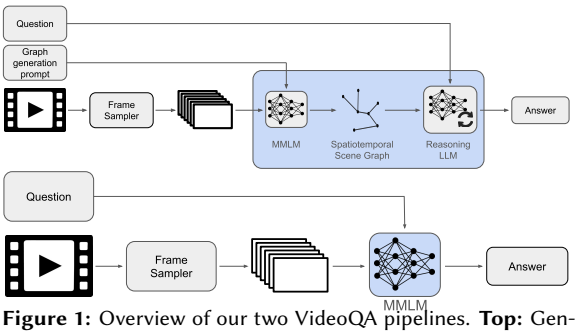

Figure 1: Overview of our two VideoQA pipelines. **Top:** Generative Graph QA: frames are first converted to scene graphs by an MLLM, aggregated, and then reasoned over by an LLM. **Bottom:** Direct VideoQA: frames are fed directly to an MLLM for VideoQA.

Crucially, this format allows for arbitrary transformations of the model responses: by aggregating nodes to form new thoughts or by looping over a node to incrementally refine existing thoughts. Video-of-thought [14] has been the first framework to apply chain-of-thought reasoning, i.e., problem decomposition and multi-hop reasoning, in the context of VideoQA tasks. While in Besta et al. [18] graphs are used to organise model responses, in Fei et al. [14] scene graphs are integrated within prompts to help solve object tracking and action analysis sub-tasks contributing to VideoQA. Inspired by these works but following a different approach, in the experiments of this paper, we inject scene graphs directly in Video-LLM prompts instructing the model to consider the graph structure for VideoQA without breaking down the video analysis in multiple (and potentially onerous) reasoning steps. Our methodological choice is also supported by recent work on image data showing that injecting scene graphs in prompts can improve the compositional reasoning abilities of Vision Language Models [17].

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.

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 [47]. 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.

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 hours. 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

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 [5], 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.

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.

Future improvements of this work also include exploring a tighter integration of the LLM component of the system with formal guarantees and structured inference, where symbolic reasoning is injected as intermediate computational layers, in the loss function [48], or by directly conditioning the graph generation. Given our focus on video input, temporal constraints could be introduced to enforce graph consistency in the STSG phase. Moreover, symbolic representations could be integrated as one additional data modality (e.g., in the form of knowledge graph embeddings), to improve the semantic consistency of graph nodes and edges, as opposed to only relying on free-form object and predicate generation via LLMs. Incorporating Neuro-symbolic inference engines into future iterations of this approach could yield significant benefits, including enhanced interpretability, formal verification of temporal coherence, and more systematic integration of domain-specific knowledge constraints for improved video understanding.

Acknowledgments

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] J. Fang, L.-l. Li, J. Zhou, J. Xiao, H. Yu, C. Lv, J. Xue, 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), 2024, pp. 22030–22040.

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

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

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

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

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

[7] 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).

[8] H. Qiu, M. Gao, L. Qian, K. Pan, Q. Yu, J. Li, W. Wang, S. Tang, Y. Zhuang, 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, 2025, pp. 3284–3294.

[9] K. Bae, J. Kim, S. Lee, S. Lee, G. Lee, 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), 2025, pp. 13744–13753.

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

[11] 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.

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

[13] P. Hitzler, A. Dalal, M. S. Mahdavinejad, S. S. Norouzi

(Eds.), Handbook on Neurosymbolic AI and Knowledge Graphs, Frontiers in Artificial Intelligence and Applications, IOS Press, 2025.

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

[15] A. Urooj, H. Kuehne, B. Wu, K. Chheu, W. Bousselham, C. Gan, N. Lobo, M. Shah, Learning situation hypergraphs for video question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14879–14889.

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

[17] C. Mitra, B. Huang, T. Darrell, 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), 2024, pp. 14420–14431.

[18] 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, 2024, pp. 17682–17690.

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

[20] A. d. Garcez, L. C. Lamb, Neurosymbolic ai: the 3rd wave, Artificial Intelligence Review 56 (2023) 12387–12406. doi:10.1007/s10462-023-10448-w.

[21] D. D. Nauck, A. Nürnberger, Neuro-fuzzy systems: A short historical review, in: Computational intelligence in intelligent data analysis, Springer, 2013, pp. 91–109.

[22] S. Aditya, Y. Yang, C. Baral, Integrating knowledge and reasoning in image understanding, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.

[23] G. Marra, S. Dumančić, R. Manhaeve, L. De Raedt, From statistical relational to neurosymbolic artificial intelligence: A survey, Artificial Intelligence 328 (2024) 104062.

[24] I. Rodin, A. Furnari, K. Min, S. Tripathi, G. M. Farinella, Action scene graphs for long-form understanding of egocentric videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18622–18632.

[25] T.-T. Nguyen, P. Nguyen, J. Cothren, A. Yilmaz, K. Luu, Hyperglm: Hypergraph for video scene graph generation and anticipation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 29150–29160.

[26] H. Zhang, X. Li, L. Bing, Video-LLaMA: An instruction-tuned audio-visual language model for video understanding, arXiv preprint arXiv:2306.02858 (2023).

[27] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, L. Yuan, Video-LLaVa: Learning united visual representation by alignment before projection, arXiv preprint arXiv:2311.10122 (2023).

[28] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).

[29] M. Maaz, H. Rasheed, S. Khan, F. S. Khan, Video-ChatGPT: Towards detailed video understanding via large vision and language models, arXiv preprint arXiv:2306.05424 (2023).

[30] J. Lei, L. Yu, T. L. Berg, M. Bansal, What is more likely to happen next? video-and-language future event prediction, arXiv preprint arXiv:2010.07999 (2020).

[31] J. Li, P. Wei, W. Han, L. Fan, Intentqa: Context-aware video intent reasoning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 11963–11974.

[32] A. Zadeh, M. Chan, P. P. Liang, E. Tong, L.-P. Morency, Social-iq: A question answering benchmark for artificial social intelligence, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8807–8817.

[33] J. Li, L. Niu, L. Zhang, From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 21273–21282.

[34] J. Xiao, X. Shang, A. Yao, T.-S. Chua, Next-qa: Next phase of question-answering to explaining temporal actions, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9777–9786.

[35] Y. Jang, Y. Song, Y. Yu, Y. Kim, G. Kim, Tgif-qa: Toward spatio-temporal reasoning in visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2758–2766.

[36] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, C. Gan, Beyond rnns: Positional self-attention with co-attention for video question answering, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 8658–8665.

[37] J. Jiang, Z. Chen, H. Lin, X. Zhao, Y. Gao, Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 11101–11108.

[38] J. Gao, R. Ge, K. Chen, R. Nevatia, Motion-appearance co-memory networks for video question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6576–6585.

[39] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, H. Huang, Heterogeneous memory enhanced multimodal attention model for video question answering, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1999–2007.

[40] F. Liu, J. Liu, W. Wang, H. Lu, Hair: Hierarchical visual-semantic relational reasoning for video question answering, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1698–1707.

[41] P. Jiang, Y. Han, Reasoning with heterogeneous graph alignment for video question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 11109–11116.

[42] Y. Li, X. Wang, J. Xiao, W. Ji, T.-S. Chua, Invariant

grounding for video question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2928–2937.
[43] J. Park, J. Lee, K. Sohn, Bridge to answer: Structure-aware graph interaction network for video question answering, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15526–15535.
[44] W. Li, H. Fan, Y. Wong, M. Kankanhalli, Y. Yang, Topa: Extending large language models for video understanding via text-only pre-alignment, Advances in Neural Information Processing Systems 37 (2024) 5697–5738.
[45] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, et al., Mvbench: A comprehensive multi-modal video understanding benchmark, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22195–22206.
[46] A. Bhattacharyya, S. Panchal, M. Lee, R. Pourreza, P. Madan, R. Memisevic, Look, remember and reason: Grounded reasoning in videos with language models, arXiv preprint arXiv:2306.17778 (2023).
[47] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 22199–22213.
[48] V. Derkinderen, R. Manhaeve, R. Adriaensen, L. Van Praet, L. De Smet, G. Marra, L. De Raedt, The deeplog neurosymbolic machine, arXiv preprint arXiv:2508.13697 (2025).