

Highest Score Award to the CVPR'2024 LOVEU Track1 Challenge: Zero-Shot Long-Form Video Understanding through Screenplay

Yongliang Wu^{1,3}, Bozheng Li^{2,3}, Jiawang Cao³, Wenbo Zhu³, Yi Lu^{3,4}
Weiheng Chi^{3,5}, Chuyun Xie³, Haolin Zheng³, Ziyue Su³, Jay Wu³, Xu Yang¹

¹Southeast University ²Zhejiang University ³Opus AI Research

⁴The University of Manchester ⁵National University of Singapore

{yongliangwu, palm_yangxu}@seu.edu.cn;

{gavin.cao, vito.zhu, sharon.xie, harley.zheng, lirian.su, jay.wu}@opus.pro;

bozhengli@zju.edu.cn; yi.lu-14@student.manchester.ac.uk; weiheng_chi@u.nus.edu

Abstract

The Long-form Video Question-Answering task requires the comprehension and analysis of extended video content to respond accurately to questions by utilizing both temporal and contextual information. In this paper, we present MM-Screenplayer, an advanced video understanding system with multi-modal perception capabilities that can convert any video into textual screenplay representations. Unlike previous storytelling methods, we organize video content into scenes as the basic unit, rather than just visually continuous shots. Additionally, we developed a “Look Back” strategy to reassess and validate uncertain information, particularly targeting breakpoint mode. MM-Screenplayer achieved highest score in the CVPR'2024 Long-form Video Understanding (LOVEU) Track 1 Challenge, with a global accuracy of 87.5% and a breakpoint accuracy of 68.8%.

1. Introduction

With the rapid development of video models, significant progress has been made in the domain of video understanding. However, the length of videos that these models can effectively handle remains limited. The long-form Video Question-Answering (LVQA) task has been introduced to explore the potential of models in understanding long-duration videos, specifically those videos longer than five minutes. The LVQA task demands a comprehensive understanding of the entire video from the global perspective, as well as the precise temporal capturing ability to answer questions about specific moments. This presents a substantially more challenging task in video understanding. Recently, the first long video understanding benchmark, MovieChat [7] has been proposed. It includes 1,000

high-quality video clips sourced from various movies and TV series, accompanied by 14,000 manual annotations. MovieChat enables quantitative evaluation of long-form video understanding capabilities in question-answering.

Most previous video models focus on end-to-end training to build a video question-answering system. Works like MovieChat [5, 7] rely on the question input to construct video representations and answers. Due to the lack of high-quality large-scale annotated data, these models have very limited capabilities in handling the LVQA task. Another line of work adopts a series of foundational models to convert video content into textual representations, which we refer to as storytelling methods [4, 10]. They leverage the powerful language understanding capabilities of LLMs to comprehend video content based on the generated scripts. However, these methods either perform captioning on each individual frame or use scene detection to segment visually continuous “shots” as the basic unit. These approaches overlook the temporal relationships between different segments, thus limiting their understanding of the video content. For example, multiple quick-cut shots in a movie often represent a single coherent event.

To address the aforementioned issues, we introduce MM-Screenplayer, an agent system endowed with multi-modal perception capabilities for tackling LVQA tasks. It comprises three essential components: the Multi-modal Perception module, which receives inputs from both visual and audio tracks of video; the Scene-Level Script Generation Module, which prompts LLMs to reassemble and comprehend shots temporally, generating high-level semantic scenes as the basic unit; and the Look back for determination Module, which extracts, analyzes, and summarizes information from video frames before and after the specified timestamp for breakpoint mode video question answering.

MM-Screenplayer achieved first place in the CVPR 2024

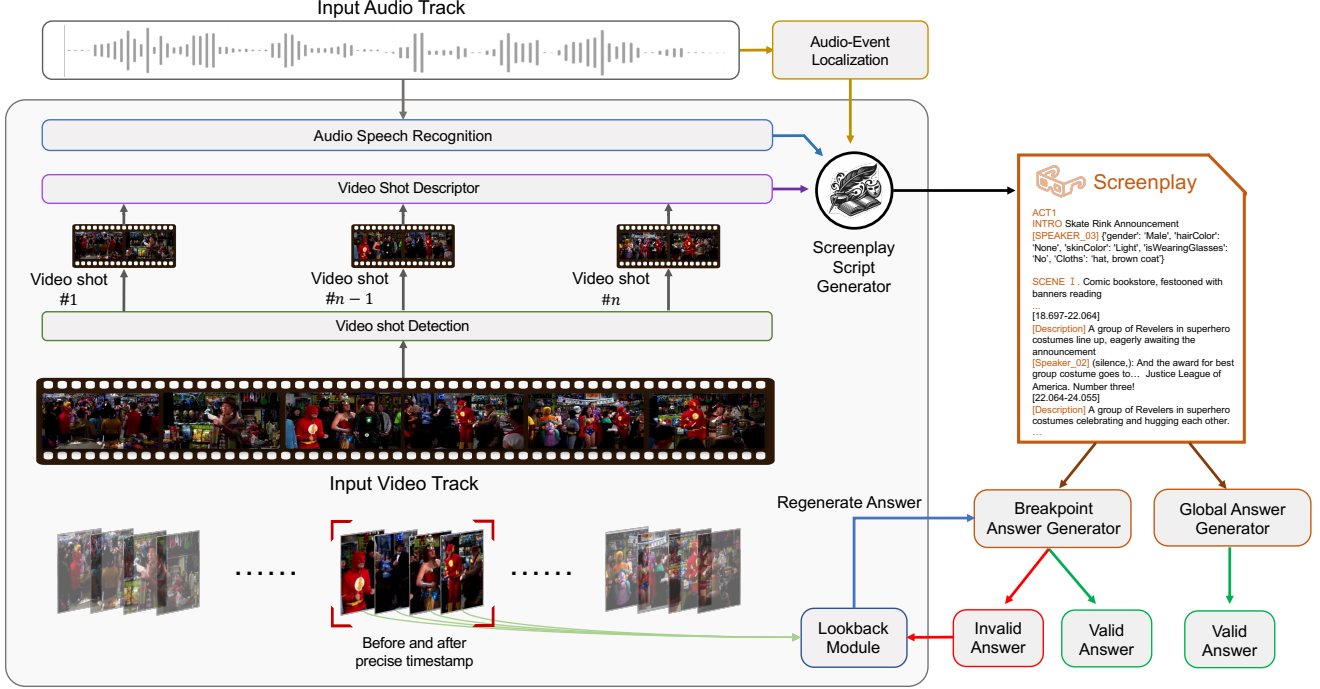


Figure 1. The overall architecture of MM-Screenplayer.

LOVEU Track 1 Challenge with a global accuracy of 87.5% and a breakpoint accuracy of 68.8% in the MoiveChat-1k long-form video understanding dataset.

2. Method

Given a video \mathcal{V} , MM-Screenplayer generates a comprehensive screenplay \mathcal{M} to thoroughly represent the content of the video. Unlike previous storytelling methods, we organize textual descriptions into higher-level semantic scenes rather than individual shots, thereby promoting a deeper comprehension of the video’s narrative. Additionally, to address certain issues of the breakpoint mode, we introduce a Look-Back mechanism: if the model is uncertain to make judgments solely based on the given screenplay, it then extracts frames to utilize extra visual information to improve the understanding of the given video, therefore producing more reliable answers. The overall architecture of our model is illustrated in Figure 1.

2.1. Multi-Modal Perception

Our model comprehensively analyzes both visual and audio tracks of a video to extract rich multi-modal information. For the visual track, we start with scene detection to divide the video into distinct shots. For each shot, frames are sampled at regular intervals. A Vision-Language Model such as GPT-4V is utilized to generate image captions for each frame [8, 9]. For the audio track, an Automatic Speech Recognition (ASR) model is applied to transcribe the di-

alogue from the audio. Additionally, we utilized an Audio Event Localization model to detect and index significant audio events throughout the video. By combining these processes, we captured a structured set of data encompassing visual shots, frame descriptions, dialogue transcripts, as well as audio events. This multi-modal extraction provides a robust foundation for advanced video analysis and understanding.

2.2. Scene-Level Scripts Generation

The concept of **scene** is fundamental in the design of screenplays, as it provides a higher-level semantic decision of the video. In contrast, simple storytelling methods merely rely on basic scene detection and treat individual shots as primary units [4]. Such an approach hinders the comprehensive understanding of the narrative. For example, in the film *Titanic* (1997), the sequence leading up to the ship’s collision with the iceberg consists of numerous quickly shifting shots. If these shots are viewed in isolation rather than treated as part of a cohesive scene, the true narrative is lost. To tackle this issue, we propose a Scene-Level Scripts Generation module that utilizes LLMs to merge shots into coherent scenes.

Initially, we arranged ASR transcripts in chronological order. For dialogues with pauses longer than 2 seconds, intervening visual content becomes significant. To address this observation, we insert a “separator” marker between sentences with gaps of more than 2 seconds, guiding the LLM to perform an initial rough split of the transcript. In

Table 1. Performance comparison on the MovieChat-1K [7] dataset against state-of-the-art methods. The best performance is highlighted in bold, while the second-best is underlined.

Method	Global		Breakpoint	
	Accuracy	Score	Accuracy	Score
<i>Evaluation via GPT-3.5</i>				
Video Chat [3]	57.8	3.00	46.1	2.29
Video LLaMA [11]	51.7	2.67	39.1	2.04
Video-ChatGPT [5]	47.6	2.55	48.0	2.45
MovieChat [7]	62.3	3.23	48.3	2.57
<i>Evaluation via Gemini Pro</i>				
MM-VID [4]	<u>58.6</u>	2.86	10.4	0.56
LLoVi [10]	58.3	<u>2.87</u>	17.8	1.03
MovieChat [7]	55.1	2.78	<u>38.5</u>	<u>1.87</u>
<i>Ours</i>	87.5	4.18	68.8	3.52

the next phase, to integrate multi-modal information and grasp the “subtext”, we insert captions of visual and audio events between these coarse split scenes. This creates a multi-modal text representation, allowing the LLM to re-process and refine the separation of the content. By analyzing visual descriptions, dialogues and audio events, the module is capable of identifying logical boundaries.

2.3. Look Back for Determination

Our pipeline employs LLMs to understand long-form video content and answer questions across different modes using tailored prompts and generated screenplays which are reusable, enhancing efficiency in LVQA by eliminating the need for re-encoding video input for each question.

A look-back mechanism is specifically designed for handling problems in breakpoint mode. We observed that in rare cases, such as the given problem requiring time localization with a higher precision, the screenplay alone may fail to extract sufficient information for answer generation.

Therefore, when the response produced by the Answer Generator was detected empty or invalid (such as: containing negative keywords such as “cannot”, “don’t know”, etc.), the model will fall back to reproduce a new answer with the incorporation of visual information.

The visual information is obtained through the video track through frame extraction. As in breakpoint mode, the exact objective timestamp is given by the question, the pipeline will extract the frames slightly before and after that timestamp consecutively, as the retrieved information is used for producing the new answer.

By combining the extracted visual time series information with the produced screenplay, the model was able to not only gather sufficient information near the provided video breakpoint but also comprehend the video plot globally, therefore being more probable to produce valid and correct answers to breakpoint problems.

Table 2. Performance Metrics with Different Components. Acronyms: SSGM - Scene-level Scripts Generation Module, LBDM - Look Back for Determination Module, G-Acc - Global Accuracy, G-Score - Global Score, B-Acc - Breakpoint Accuracy, B-Score - Breakpoint Score.

SSGM	LBDM	G-Acc	G-Score	B-Acc	B-Score
×	×	66.7	3.60	48.5	2.51
✓	×	85.6	4.18	54.8	2.77
✓	✓	87.5	4.18	68.8	3.52

3. Experiments

3.1. Experimental Settings

The LVQA Challenge offers 170 videos as the test set to evaluate the model’s performance. The evaluation process is processed by the competition organizer using unreleased answers. In our proposed screenplay generation pipeline, we employed GPT-4-turbo [1] as the LLM driving all text script processing tasks. For visual description generation, we cherry-picked GPT-4o as the corresponding VLM. Additionally, we integrated whisperX [2] as the ASR model. Gemini-1.5 pro [6] was chosen as the audio analyzer for audio event localization. The versions and parameters of the LLM and VLM are fixed to ensure reproducibility. All experiments are conducted on a single T4 GPU without any extra training process.

3.2. Main Results

As shown in Table 1, MM-Screenplayer achieved top performance on the MovieChat-1K test set, with a global accuracy of 87.5% and a global score of 4.18. In breakpoint mode, our method attained an accuracy of 68.8% and a score of 3.52, ranking highest among all participants.

The outstanding performance in both global and breakpoint modes demonstrates that our proposed screenplay format plays a pivotal role in representing long video content and providing rich contextual information. This enables LLM to understand long-form video content and accurately answer questions. Compared to the locally reproduced MovieChat [7] baseline, our pipeline generates answers that exhibit a comprehensive understanding of long-form video content and accurate temporal capture ability.

3.3. Ablation Study

We conducted extensive ablation studies on our proposed Scene-Level Scripts Generation Module and Look-Back for Determination Module as shown in Table 3.2. The results show that organizing shots into higher-level scenes as the basic unit significantly improves global accuracy. Furthermore, the introduction of the Look-Back strategy greatly enhances performance in breakpoint mode. These findings

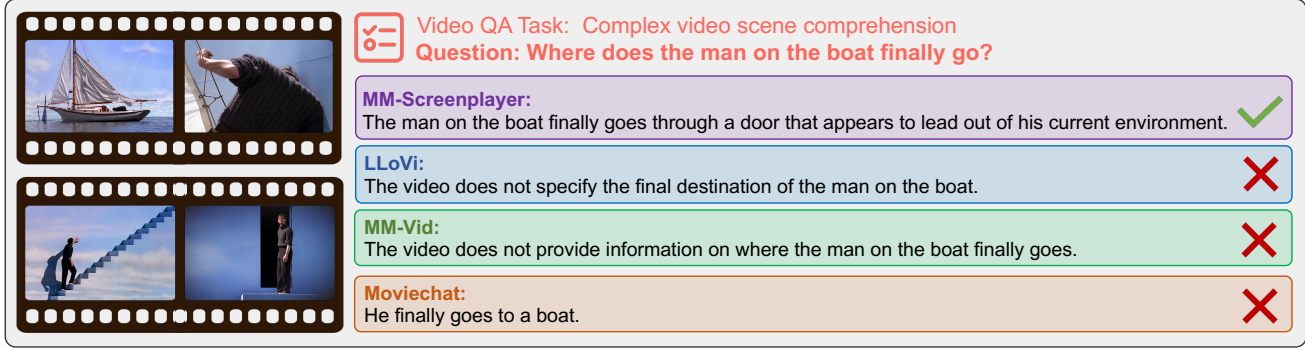


Figure 2. Comparison of answers produced by MM-Screenplayer and other state-of-the-art methods for a question from MovieChat1K-testset. Our method produced significantly better answers while all other methods’ answers were incorrect.

demonstrate the effectiveness of both modules. However, the results also indicate that solely relying on screenplay content is unreliable for addressing some detailed issues in breakpoint mode. This is because our visual descriptions might not have adequately captured the necessary information.

3.4. Qualitative Results

Figure 2 presents one of our answers on MovieChat-1K test set. MM-Screenplayer produced significantly better answers which precisely captured the environmental transition of this clip of “Truman’s World”, while other methods either produced incorrect destinations or could not answer the question at all. Screenplay enables our method to accurately capture the high-level global semantics of this movie clip while the look-back mechanism could further guarantee the understanding of specific moments along with the context provided by the screenplay.

4. Conclusion

In this paper, we introduced MM-Screenplayer, a comprehensive video understanding system with multi-modal perception capabilities that can convert videos of arbitrary length into higher-level textual representations. Our innovative approach organizes video content into scenes as the basic unit and employs a “Look Back” strategy to reassess and validate uncertain information. MM-Screenplayer achieved the highest score in the CVPR’2024 LOnG-form VidEo Understanding (LOVEU) Track 1 Challenge, demonstrating exceptional proficiency in long-form video understanding. This accomplishment underscores the effectiveness of our method in comprehending and analyzing extended video content for accurate question-answering.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [2] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023. 3
- [3] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3
- [4] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*, 2023. 1, 2, 3
- [5] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 3
- [6] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3
- [7] Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 1, 3
- [8] Yongliang Wu and Xu Yang. A glance at in-context learning. *Frontiers of Computer Science*, 18(5):185347, 2024. 2
- [9] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [10] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023. 1, 3
- [11] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, 2023. 3