CharacterQA: A Corpus for Multimodal Character Conversational Movie Question Answering

Anonymous ACL submission

Abstract

The rapid advancement of Large Language Models has sparked extensive exploration of their applications across various fields. Among them, the personalized conversation based on characters in movies is an attractive research area. To achieve such comprehensive conversations, the integration of extensive multimodal information, notably visual content alongside textual data, is crucial. This necessity underlines the significance of multimodal insights for enriching personalized conversations, thereby further emphasizing the urgent need for a sophisticated multimodal character conversational dataset. To this end, we introduce CharacterQA, a novel video question-answering (QA) dataset for multimodal character conversation in movies. The dataset consists of 101 selected Chinese movies, each of which is annotated with the main character profiles, the character information of the scripted conversations and their timestamps. Furthermore, a set of questions from various designed tasks and their detailed answers are annotated. Most of those questions require taking into account visual signals for logical comprehension of movie characters and plots. Subsequently, we adopt an advanced multimodal large language model MovieGPT to evaluate the CharacterQA dataset. The results yield insightful findings that are expected to drive further development of multimodal large language models in the character conversation field.

1 Introduction

In the past few years, the development of social media has greatly contributed to the user demand for personalized character conversation, and consequently inspired significant attention from researchers. Recently, breakthroughs in pre-trained large language models (LLMs) have led to a paradigm shift in the natural language processing community, which brings novel challenges for character conversation (Brown et al., 2020; OpenAI, 2023a; Touvron et al., 2023a; Liu et al., 2023b; Peng et al., 2023). Currently, personalized character conversation mainly focuses on text-only domains (e.g., character.ai) (Shao et al., 2023; Park et al., 2023), largely overlooking diverse multimodal applications in real-world scenarios.

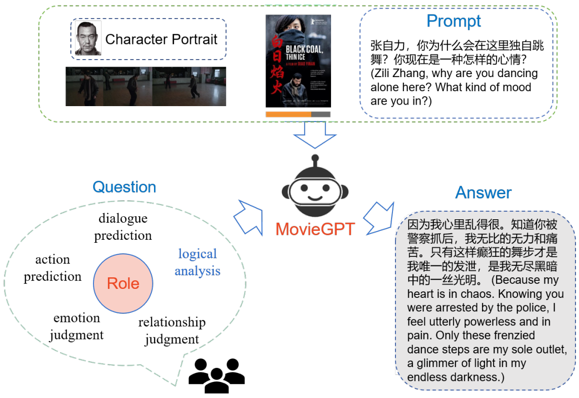


Figure 1: An example of multimodal character QA. Our CharacterQA dataset enables MovieGPT to perform character conversations by watching movies and respond to personalized questions in different scenes.

As shown in Figure 1, given a video taken from a movie, a user may wish to play the character "Zhizhen Wu" and ask the character "Zili Zhang": "Zili Zhang, why are you dancing alone here?". If only the textual information is available, the model is unable to answer such a question since capturing the nuances of expressions and body language during dance requires understanding visual content. Conversely, compared to text-only settings, multimodal character conversations offer enhanced vividness and practicality with the visual content, facilitating much easier user interactions. Regrettably, prevailing datasets fail to equip models with such character conversation capabilities.

In light of the above considerations, we construct the first multimodal character conversation dataset CharacterQA in this paper. Diverging from existing multimodal conversation datasets primarily based on pure English language (Tapaswi et al., 2016; Lei et al., 2018; Castro et al., 2022; Xiao et al., 2021), this dataset is derived from 101 Chinese films obtained from online platforms. Specifically, we annotate attributes such as the names, roles, and personalities of the main characters in each film, as well as their dialogues with corresponding timestamps. To evaluate character conversation ability, five tasks of varying difficulty are designed, including dialogue prediction, action prediction, relationship judgment, sentiment analysis, and logical analysis. Among them, the sentiment analysis is multiple choice questions, while the others are open-ended questions. Some questions are relatively straightforward as the answers can be found within the dialogues, while others pose greater challenges. These challenging questions require a deep understanding of the movie content, character profiles, and long conversation contexts. This depth of comprehension is necessary to capture the nuances and unique styles of the characters' language. Furthermore, certain questions even require the ability to reason across the dialogues and the movie content based on broader commonsense knowledge related to the question.

We also developed a multimodal LLM called MovieGPT, and evaluated it alongside various LLMs using our CharacterQA dataset to evaluate their character conversation abilities. Thorough analysis of the results indicates that the CharacterQA dataset poses significant challenges for multimodal character conversation, and existing LLMs are insufficient for character portraits and visual understanding in multimodal scenarios.

This study makes several contributions. Firstly, we present a novel dataset CharacterQA for Chinese multimodal character conversation, comprising five distinct designed tasks that emphasize the understanding of character traits and multimodal content. Secondly, we introduce a multimodal pretrained LLM MovieGPT tailored for character conversation. Thirdly, we conduct extensive evaluations on CharacterQA with MovieGPT and several mainstream LLMs, emphasizing the challenges inherent in the multimodal character conversation.

2 Related Work

Multimodal Conversational LLMs and Character-play Datasets. The success of LLMs has catalyzed advancement in multimodal conversational LLMs, such as Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023a), MiniGPT-4 (Zhu et al., 2023), BLIP2 (Li et al., 2023b), and mPLUG-Owl (Ye et al., 2023d). These methods have extensively explored the visual encoders and training strategies of multimodal LLMs. However, they were not originally tailored for character-play scenarios, and previous evaluations reveal a deficiency in their capacity for robust character-playing (Shen et al., 2023; Huang et al., 2023a; Wang et al., 2023b). Concurrently, although the potential for character-playing within the LLMs has been acknowledged, the existing character-playing datasets are limited to text-only formats, lack multimodal annotations, and feature a small number of characters (often less than 200) (Tu et al., 2024; Chen et al., 2023b). These limitations highlight the importance of our proposed CharacterQA, which aims to enhance training and evaluation for character-play capabilities.

Multimodal QA Datasets. Several datasets focusing on multimodal video QA have been developed, including MovieQA (Tapaswi et al., 2016), VideoQA (Zhu et al., 2017), TV-QA (Lin et al., 2023), Life-QA (Castro et al., 2020), NexT-QA (Xiao et al., 2021), and Wild-QA (Castro et al., 2022). As shown in Table. 1, the existing datasets primarily focus on the visual comprehension capabilities of models, lacking the necessary annotations of the intrinsic profiles of characters essential for multimodal character conversation. Consequently, achieving an effective evaluation of this task becomes challenging. Focusing on multimodal character-based conversation, our dataset includes detailed character profiles, manually curated conversational texts, complete sets of movies, and annotations for five distinct character-based conversation tasks, as described in Section 3.

3 CharacterQA Dataset

3.1 Dataset Summary

Our CharacterQA dataset comprises 101 carefully selected Chinese movies, with an average duration of 102 minutes per movie. Among these movies, 90 are dubbed in Standard Mandarin, while 11 are dubbed in various regional Chinese dialects. The selected movies span a range of release dates, from as early as 1984 to the most recent in 2023. As shown in Figure. 2, each movie contains an average of 405 lines of dialogue, with each line annotated with its timestamp and the corresponding character. To accommodate the character conversation task,

Dataset	Lang.	Domain	Annotation	QA Type	Role Inf	Video#	QA#	Dur.(s)
MovieQA	En	Movie	Man	MC	N	6.7k	6.4k	203
VideoQA	En	Web	Aut	MC	N	109k	390k	33
TVQA	En	TV Shows	Man	MC	N	21.8k	152k	76
LifeQA	En	Daily Activaty	Man	MC	N	0.3k	2.3k	74
NExTQA	En	Daily Activity	Man	OE+MC	N	5.4k	52k	44
WildQA	En	Wild Activity	Man	OE	N	0.4k	0.9k	71
Ours	Ch	Movie	Aut+Man	OE+MC	Y	101	25k	6024

Table 1: Comparison between our dataset and representative existing datasets for videoQA. "Lang." denotes the language of the data, "En" for English, and "Ch" for Chinese. "Annotation" indicates whether the data is annotated manually or automatically. "Aut" stands for automatical and "Man" stands for manual. "QA Type" denotes whether the answers are multiple-choice (MC) or open-ended (OE). "Dur. (s)" is the average duration of the videos in seconds. "Role Inf" is the character information.

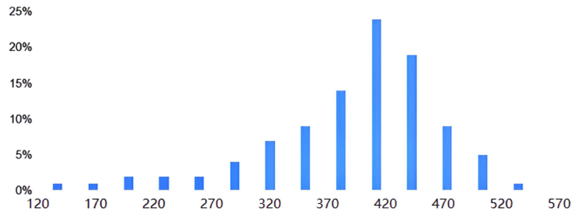


Figure 2: Illustration of dialogue counts for each movie.

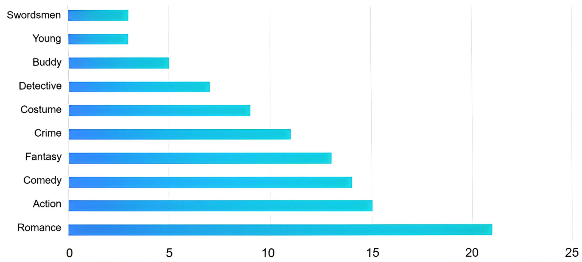


Figure 3: Illustration of genres for the selected 101 movies.

we annotate detailed character profiles delineating the names, personalities, and backgrounds of 3 to 5 main characters within each movie. Each profile is carefully crafted to summarize and portray their distinctive characteristics in about 70 words of Chinese description.

Based on the detailed profiles and the labeled dialogues, five different video question-answering tasks are designed, including dialogue prediction, action prediction, relationship judgment, sentiment analysis, and logical analysis. Particularly, there are 25,000 QA pairs in total for those five tasks, with 5,000 pairs per task evenly distributed across the 101 movies. Note that the selected movie clips used for task annotations are evenly distributed throughout the entire movie, facilitating subsequent model training and validation. Each QA instance corresponds to a one-minute selected movie clip, the plot of which is consecutive to avoid escalating the difficulty of video comprehension.

3.2 Dataset Details

As illustrated in Figure. 3, to ensure the diversity of the dataset, we strive to gather a wide spectrum of Chinese movies spanning across different genres, including romance, action, comedy, fantasy, etc.. Note that a portion of our dataset is selected from the Movie101 dataset (Yue et al., 2023). In order to obtain dialogue transcripts that match the characters and timestamps in the movies, we use the "Tongyi Tingwu" software of Alibaba for audio-to-text transcription, which generate dialogue corresponding to timestamps. After minimal manual adjustments, we acquire the necessary dialogue content. This method forms a robust foundation for constructing the dataset of character conversation. More details of the movie and main character selection can be found in Sec. A.1.2 and Sec. A.1.3.

The second stage of dataset construction involves character profiles and five different types of VideoQA tasks. Table. 2 shows the QA examples for each task. All annotations are carried out on an enterprise crowdsourcing platform. All workers are proficient in Chinese, possess a solid educational background to ensure accurate comprehension of video content, and produce corresponding descriptions precisely. Note that workers need to have completed at least 100 prior tasks on the platform with a minimum approval rate of 95%. Additionally, we conduct daily spot checks on annotations written by each worker to verify their relevance to

Task	Evaluation	QA Examples（movie: Goodbye Mr.Loser）
Dialogue prediction	BLEU,CIDEr, ROUGE-L, GPT,Human	start_time: 01:26:12 end_time: 01:26:42 Q:夏洛接下来要说什么话？(What will Xia Luo say next?) A:可我最心爱的女人被别人抢走了。(But the woman I love the most is taken away from me.)
Action prediction	BLEU,CIDEr, ROUGE-L, GPT,Human	start_time: 01:01:40 end_time: 01:02:10 Q:袁华接下来会做出怎么样的行为？(What will Yuan Hua do next?) A:袁华接下来会在漫天飞雪的环境下哭泣。(Yuan Hua will cry in the midst of falling snow.)
Relationship judgment	BLEU,CIDEr, ROUGE-L,GPT, Human,Accuracy	start_time: 00:08:28 end_time: 00:09:10 Q:夏洛和马冬梅什么关系？(What's the relationship between Xia Luo and Ma Dongmei?) A:夏洛和马冬梅是夫妻关系。(Xia Luo and Ma Dongmei are husband and wife.)
Sentiment analysis	Accuracy	start_time: 01:01:40 end_time: 01:02:10 Q:袁华此时的心情如何？(What is Yuan Hua's mood at this time?) A:此时袁华的情绪是悲伤的。(At this time Yuan Hua's mood is sorrow.)
Logical analysis	BLEU,CIDEr, ROUGE-L, GPT,Human	start_time: 01:07:00 end_time: 01:08:10 Q:袁华此时作诗和之前作诗时的差别在哪？分析原因。 (What are the differences in Yuan Hua's poetry now compared to before? Analyze reasons.) A:袁华之前作诗带有批判性，让夏洛很没有面子，现在作诗则极尽谄媚，巴结夏洛。 原因在于夏洛和袁华的社会地位发生了翻天覆地的变化，袁华现在穷困潦倒，不比之前。 (Previously,Yuan Hua's poetry bore a critical tone, which led to Xia Luo losing face. However,there has been a significant shift in his poetic style, now excessively fawning. This change can be attributed to the stark reversal in social status between Xia Luo and Yuan Hua.)

Table 2: The evaluation methods and QA examples for different tasks varied

	Optional items						
Emotion	愤怒 anger	厌恶 disgust	惊讶 amazed	悲伤 sorrow	喜悦 joyful	中立 neutrality	恐惧 fear

Table 3: The seven options for sentiment analysis

Character portrait	Name	Identity(translation)
夏洛特烦恼 Goodbye Mr.Loser	夏洛 Xia Luo	Xia Luo is a vengeful, greedy, and timid person who, in pursuit of vanity, puts on airs. However, at the same time, he is someone who has always harbored dreams of music.
夏洛特烦恼 Goodbye Mr.Loser	马冬梅 Ma Dongmei	Ma Dongmei has a straightforward and somewhat tomboyish personality, lacking a bit of feminine charm. She is unburdened by trivialities, upright, and stands up for justice. Ma Dongmei is a simple, dedicated, hardworking, and capable individual.
前任3-再见前任 The Return of the Exes	孟云 Meng Yun	Meng Yun is a career-oriented man with ambitious goals in his professional life, displaying a proactive and upwardly mobile attitude. Although he appears composed on the surface, there are fluctuations in his inner world.
我不是药神 Dying to suivive	程勇 Cheng Yong	Cheng Yong is a small-time merchant who peddles Indian God Oil. He is opportunistic, selfish, engages in domestic violence against his wife, and is generally self-centered, bullying the weak and fearing the strong. However, later on, he rediscovers his inherent kindness, starts helping others, and shows a sense of responsibility.

Table 4: The names and identity descriptions of characters from several movies are displayed

the respective videos. We require workers to first watch the selected movie, describing the main characters' personalities using keywords and sentences, and providing concise descriptions of character profiles. Each character profile includes the name, individuality, and identity of a movie character, as shown in Table. 4. When selecting suitable movie clips, we annotate the start and end times, with the end time being one second before the answer appears. The annotation process for the dataset spans three months, involving 39 qualified workers who contribute annotations for 25,000 questions across 101 movies. Additionally, corrections are made to 40,905 dialogue data entries, and descriptions for 398 character profiles are provided. More details of the quality control of annotations is depicted in Sec.A.1.4. The design details of the five VQA tasks are as follows:

- **Dialogue prediction.** The prompt format is "What will [movie character] say next?" and the answer is the next line of dialogue for the character in the movie.
- **Action prediction.** The prompt format is "What will [movie character] do next?" and the answer is the next action or expression of the character in the movie.
- **Relationship judgment.** The prompt format is "What is the relationship between [movie character A] and [movie character B]?" and the answer is the relationship between the characters in the specific scene of movies.

- **Sentiment analysis.** The prompt format is "How is [movie character] feeling at this moment?" and the answer describes the emotion of a character based on movie clips. The response should be chosen from the following seven emotions: anger, disgust, joy, sorrow, neutral, surprise, and fear, as shown in Table. 3.

- **Logical analysis.** The prompt format is: "Why is [movie character] engaging in a specific behavior, expression, or action?" or "[movie character] changes behavior from previous action to current action, analyze the reasons." Responses must be based on the current movie clip and long context, providing explanations.

We choose three native Chinese speakers to cross-validate the modified script dialogues. They verify the accuracy of the dialogue content by watching 101 Chinese movies, first confirming the correctness of the dialogue and then checking the alignment of the dialogue with timestamps and the characters in the movies. Corrections are made for any inconsistencies. For character profiling validation, we randomly assign 398 selected main movie characters to 50 individuals, with each character assessed by two people familiar with Chinese and relevant movies. They provide ratings for the character profiles, and a consensus with satisfaction levels exceeding 85% is considered a pass; otherwise, it is reassessed by annotators. The validation for the five video QA tasks is relatively straightforward. We randomly reassign the annotated QA pairs to two additional individuals, who then assess whether the answers are consistent with the movie plot and characters. In cases of inconsistency, modifications are made.

3.3 Comparison with VideoQA Datasets

As summarized in Table. 1, most existing videoQA datasets focus on English question-answering, with a primary emphasis on visual understanding. The QA pairs typically interpret content from specific excerpts. In contrast, our dataset is dedicated to conversations between characters in a multimodal context, specifically within the context of movie scenes. As a Chinese QA dataset, our questions are designed to revolve more around the characters in movies, combining video and long background context to predict character actions or dialogues. Besides, our dataset is compatible with both open-domain and closed-domain QA tasks. It comprises video clips and textual descriptions that are significantly longer than those in existing video narration datasets. The average length of video clips is 6024 seconds, surpassing the average length of current video datasets by a considerable margin.

4 MovieGPT Model

4.1 Design Principles

As illustrated in Figure. 4, Our MovieGPT model is a multimodal large language model based on the transformer architecture (Vaswani et al., 2017). The model utilizes the pre-trained Bloom-7B as its backbone (Workshop et al., 2022). Besides, the visual module contains a ViT-L/14 (Dosovitskiy et al., 2021) visual encoder and a connecting layer visual abstractor (Liang et al., 2022) .

To enable the model to engage in character conversation and question-answering within specific scenarios, we facilitate the learning of character-specific knowledge and memories by providing an abundance of dialogues from various movie characters. Additionally, detailed descriptions of character individuality and identity are provided to enhance the model's understanding of the characters' language styles. Given video clips from movies, we task the model with comprehending the characters in the film and engaging in conversation or answering questions in specific movie scenes. The character responses generated by the model should adhere to several criteria of faithful representation, including: (1) Lexical Consistency – the model should reflect the personality of a character, ensuring consistency with the character's unique conversation style; (2) Dialog Authenticity – the generated responses should not only be contextually relevant but also align with the content and plot of the movie.

4.2 Model Implementation

At the current stage, existing multimodal LLMs can effectively process visual information. However, they have certain limitations in understanding character conversation, particularly in Chinese conversational contexts (Yang et al., 2022; Ye et al., 2023b; Zhang et al., 2023; Zheng et al., 2023). Our exploration focuses on the understanding of Chinese character conversations in multimodal scenarios using LLMs. More training details are depicted in Sec. A.2.1. To facilitate the multimodal character

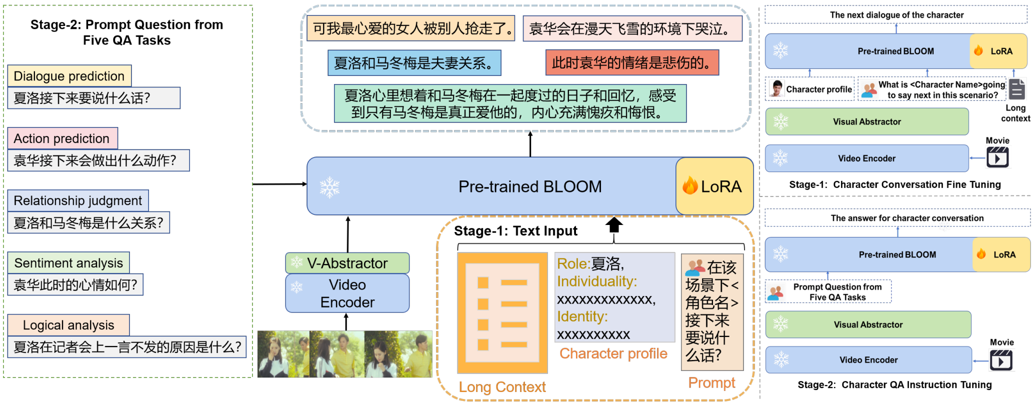


Figure 4: Illustration of the two-stage training of our MovieGPT. In the first stage, the model is fed with long context from the characters, the movie clip, and character profiles. Given the fixed prompt, the goal is to predict the next dialogue of the character. The training dataset of the second training stage is the labeled QA pairs of five designed tasks, namely "dialogue prediction, action prediction, relationship judgment, sentiment analysis, and logical analysis", combined with the movie clips.

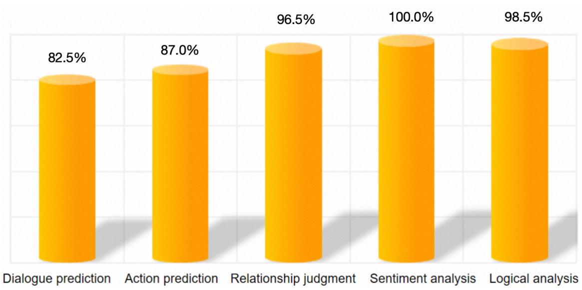


Figure 5: The accuracy of human testing is evaluated across five tasks, with 200 samples extracted for testing from a pool of 25k QA pairs for each task.

QA, our MovieGPT is trained with the following stages.

Character conversation fine-tuning At this initial stage, the pre-trained Bloom and the visual module remains in a frozen state. The model undergoes training with LoRA fine-tuning (Hu et al., 2021), with inputs comprising long context, movie clips, character profiles, and prompts in a standardized format. The expected output is the next line of dialogue of the given character.

Character QA instruction tuning. During this second stage, the parameters of Bloom-7B and visual module are still frozen. The MovieGPT is trained with LoRA fine-tuning, with inputs comprising the prompt question of five character QA tasks along with the corresponding movie clip. The expected output is the response of the character to the question.

5 EXPERIMENTS

In this section, our CharacterQA is evaluated by MovieGPT and several mainstream LLMs. Through experiments, we investigate whether multimodal LLMs can be trained efficiently with the dataset to achieve character-based contextual understanding. Furthermore, we explore the character conversation capability of the model to characterize and interact with users across five challenging tasks, including dialogue prediction (DP), action prediction (AP), relationship judgment (RJ), sentiment analysis (SA), and logical analysis (LA).

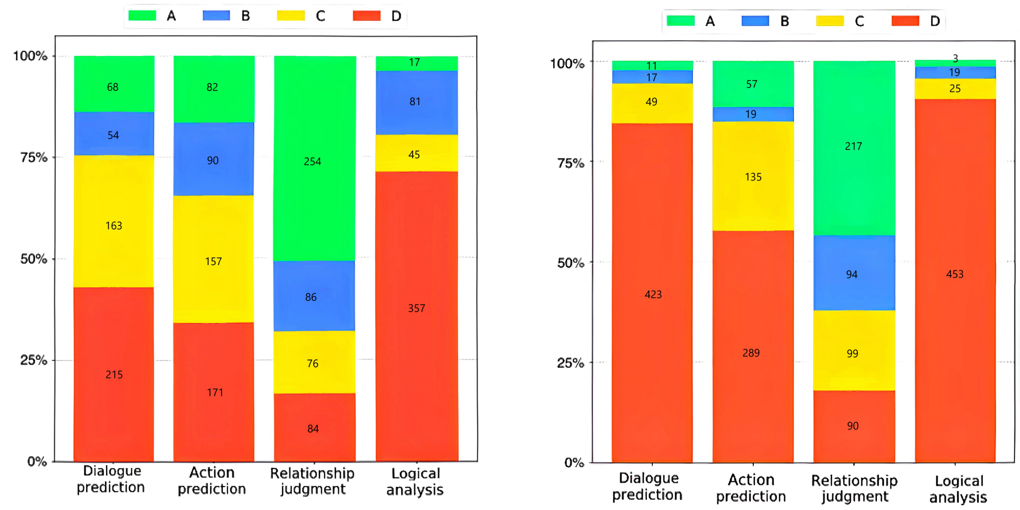
5.1 Experimental Setup

The CharacterQA dataset is uniformly sampled to construct the training, validating, and testing sets, with 20,000, 2,500, and 2,500 QA instances, respectively. Note that the long context for each question is conversations of 30 minutes before the movie clip. To ensure efficient QA training, we only collect movie clips as video inputs, which are limited to one minute. This is because existing models encounter difficulties in encoding long videos.

5.2 Task Evaluation

To demonstrate the practicality of five tasks in real-world scenarios, we conduct manual testing with 200 randomly selected samples for each task. Specifically, the input of the model is provided to human participants for responses, which are then assessed by three native Chinese speakers. The human testing results of five tasks are in Figure. 5.

To comprehensively verify the performance of character conversations, various evaluation metrics are adopted on the five VideoQA tasks, and the corresponding results are presented in Table. 5. Besides, we invite five native Chinese speakers



(a) Human evalutaion

(b) GPT-4 evalutaion

Figure 6: Human and GPT-4 evaluation of MovieGPT's performance on four open-domain QA tasks. The scoring is done on a scale of "very consistent (A)", "somewhat consistent (B)", "fairly consistent (C)", and "not consistent (D)".

Task	BLEU-4	CIDEr	ROUGE-L	Accuracy
Dialogue prediction	15.28	21.89	7.67	-
Action prediction	29.21	41.02	0.75	-
Relationship judgment	73.62	78.30	33.89	45.76%
Sentiment analysis	-	-	-	37.29%
Logical analysis	4.45	19.51	0.00	-

Table 5: Evaluation results of our MovieGPT on five tasks with several different metrics. For sentiment analysis, only Accuracy is calculated.

to watch those movies and collectively evaluate whether the generated answers of the model align with the standard ones. Ratings are assigned across four levels: very consistent (A), somewhat consistent (B), fairly consistent (C), and not consistent (D). As shown in Figure. 6, we also employ GPT-4 to assess MovieGPT's responses based on alignment scores (OpenAI, 2023a). Particularly, we provide the long context of dialogue and the character profile as the prompt, enabling GPT-4 to score responses of our MovieGPT on different tasks.

The results in those figures and tables illustrate that existing multimodal LLM architectures still face great challenges in five character conversational videoQA tasks. Except for sentiment analysis which is in a multi-choice form, other four tasks are open-domain questions. The results of these open-domain tasks, except for relationship judgment, deviated significantly from expectations, especially in the dialogue prediction and logical analysis tasks. This is mainly because character dialogues and storylines in movies are full of drama and discontinuity, while existing models are unable to realize movie story reconstruction and reasoning through simple visual encoding. Notably, in comparison to human evaluations, the evaluations based on GPT-4 tend to assign lower scores to model responses. This discrepancy arises because human evaluators effectively take into account the content of the movie plot. For responses to various open-domain questions, human evaluators, even when there is some deviation from the standard answer, assign higher scores as long as the responses align with the movie context. In contrast, GPT-4 places a direct emphasis on the alignment between model responses and standard answers.

5.3 Comparison Experiments

As demonstrated in Table. 6, we compare MovieGPT with several state-of-the-art LLMs (Bard (Thoppilan et al., 2022), Claude (Bai et al., 2022), GPT3.5 (Brown et al., 2020) and GPT-4) and two multimodal LLMs (NExT-GPT (Wu et al., 2023) and Video-LLaVA (Lin et al., 2023)) on five tasks to verify its character conversational abilities. Due to the lack of visual input capability in text-only LLMs, character dialogue from relevant movie clips serves as a proxy for video data. These baselines forego fine-tuning for specific tasks, opting instead for in-context learning (Li et al., 2023a). For two multimodal LLMs, they undergo the same

Model	Dialogue prediction	Action prediction	Relationship judgment	Sentiment analysis	Logical analysis
Bard	0.35	7.29	36.27	17.00%	4.21
Claude-2	1.43	14.34	63.32	30.50%	9.43
GPT-3.5	0.94	11.84	76.57	28.50%	12.04
GPT-4	2.74	21.02	**87.65**	35.50%	**16.87**
NExT-GPT	4.67	11.64	57.98	23.64%	2.31
Video-LLaVA	6.44	17.63	61.04	31.86%	3.37
MovieGPT	**15.28**	**29.21**	73.62	**37.29%**	4.45

Table 6: Comparison of our method with Bard, Claude-2, GPT-3.5, and GPT-4 on five tasks, namely "dialogue prediction, action prediction, relationship judgment, and logical analysis" using the BLEU-4 metric, and "sentiment analysis" using accuracy as the measure.

Model	DP	AP	RJ	SJ	LA
w/o movie clips	13.13	26.43	70.44	31.18%	3.32
w/o role profile	11.46	28.13	66.21	36.91%	2.05
w/o long context	3.47	8.25	54.23	21.66%	3.87
MovieGPT	15.28	29.21	73.62	37.29%	4.45

Table 7: Our method in comparison with others through ablation experiments under different scenarios.

two-stage training as our MovieGPT. The results show that, although MovieGPT may not exhibit comparable performance to GPT-4 in relationship judgment and logical analysis, it outperforms all baselines over the other three tasks by a large margin, highlighting the importance of the understanding of visual contents. More detailed comparisons can be found in Sec. A.2.3.

5.4 Ablation Studies

The results of three ablation variants are shown in Table. 7, where the movie clips, character profile, and long context input are removed, respectively. When the movie clips are missing, we replace them with the corresponding character dialogue. More ablation results are shown in Sec. A.2.4. It can be observed that the model obtains substantial improvements across all tasks even if only the brief movie clips are adopted, which demonstrates the important role of multimodal video semantics in the character conversation. Moreover, the absence of the long context greatly impacts the model performance, especially in dialogue prediction. This is reasonable since the long context is important for understanding character backgrounds and their expression habits. Comparatively, the influence of removing character profiles is more pronounced in logical analysis and relationship judgment, which rely more on specific information such as character identity and individuality.

6 Conclusion

We propose a Chinese multimodal character question answering dataset, comprising 101 carefully selected Chinese movies. Compared to existing datasets, our CharacterQA focuses on personalized comprehension in the Chinese multimodal conversational settings. In addition to specially annotated script conversations and character profiles, we design five videoQA tasks to evaluate character QA abilities. In addition, we built a multimodal character QA model called MovieGPT, and conducted various experiments to evaluate the multimodal character conversation capability of mainstream LLMs. The results demonstrate that the character-based QA tasks are still very challenging for current models. This requires exploring long-distance visual semantics, and mining character personality profiles needed for personalized reasoning. It also involves extending our CharacterQA to a broader range of languages and more complex problems, thereby indicating our future research directions.

7 Limitations

In this study, we investigate a multimodal LLM for character conversation through movie watching, a challenging task that requires ongoing efforts. Our work remains limited in several aspects: firstly, data constraints exist as our selected movies and character roles are limited, which are insufficient to encompass the diverse landscape of existing Chinese films. Future endeavors could benefit from a broader selection. Secondly, the foundational model—results from supervised fine-tuning—are highly influenced by factors such as the pre-training data distribution, model architecture, and scale. Subsequent work may explore trainable agents based on more potent and LLMs.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. 2020. Lifeqa: A real-life dataset for video question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4352–4358.

Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai Burzo, and Rada Mihalcea. 2022. Wildqa: In-the-wild video question answering. *arXiv preprint arXiv:2209.06650*.

Ke Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *ArXiv*, abs/2306.15195.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023b. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, W. Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Jiao Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *ArXiv*, abs/2304.15010.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023a. Chat-gpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023b. Language is not all you need: Aligning perception with language models. *ArXiv*, abs/2302.14045.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916.

OpenAI. 2023a. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2023b. Gpt-4v(ision) system card.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023a. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *CoRR*, abs/2307.02499.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023b. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023c. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023d. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. 2023. Movie101: A new movie understanding benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4669–4684, Toronto, Canada. Association for Computational Linguistics.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

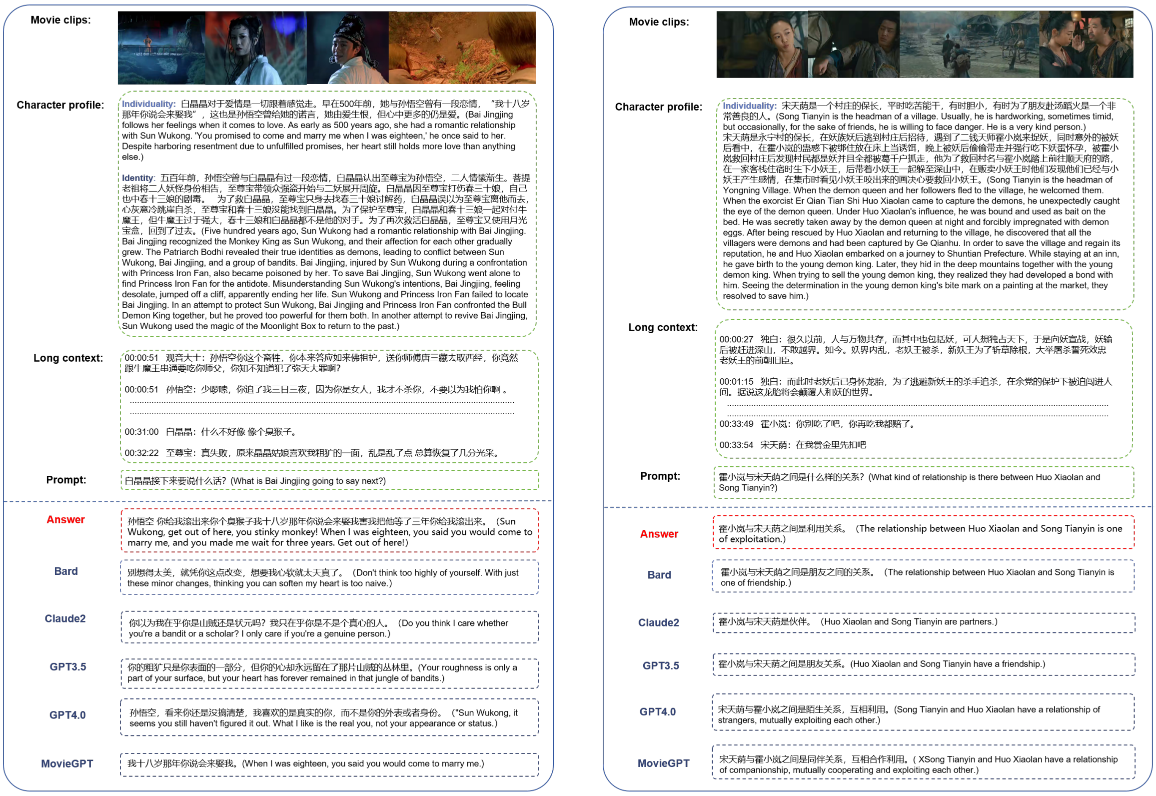
Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592.

Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124:409–421.

A Appendix

A.1 Dataset Details

A.1.1 Visualization of our CharacterQA



(a) An example of the dialogue prediction task. (b) An example of the relationship judgment task.

Figure 7: Examples of the dialogue prediction and relationship judgment task. Given the labeled movie clip, character profile and long context, the answers of different LLMs to the question prompt are also illustrated.

In this section, we provide visualization results for the CharacterQA dataset. As shown in Figure. 7, Figure. 8, and Figure. 9, each of the five tasks of dialogue prediction, action prediction, relationship judgment, sentiment analysis, and logical analysis demonstrates a character QA sample, as well as the answers of our MovieGPT and other different LLMs like Bard, Claude-2, GPT-3.5, GPT-4, to the relevant multimodal question.

As illustrated in Figure 7(a), the task of dialogue prediction poses a formidable challenge. This is primarily due to the response " 孙悟空 你给我滚出来你个臭猴子我十八岁那年你说会来娶我害我把他等了三年你给我滚出来。 Sun Wukong, get out of here, you stinky monkey! When I was eighteen, you said you would come to marry me, and you made me wait for three years. Get out of here!" being imbued with intense personal emotion. Moreover, compared to preceding dialogues, this response appears particularly abrupt and necessitates a comprehensive understanding through the integration of the movie clip for an appropriate answer. Consequently, the responses generated by the majority of LLMs are overly subdued, lacked the character's emotional coloring, and strayed far from the answer. Only GPT-4's answer was close to the edge, and only our movieGPT's answer was very close to the edge. This also shows that the two-stage training is very effective.

As for the relationship judgment task shown in Figure 7(b), if one merely floats on the understanding of textual content, the answer "霍小岚与宋天荫之间是利用关系。 (The relationship between Huo Xiaolan and Song Tianyin is one of exploitation.)" will be difficult to obtain (It can be seen that Bard, Claude2, GPT3.5, all these text-only LLMs answered poorly, and only by combining the visual information, GPT-4 and MovieGPT can answer the question accurately).

(a) An example of the action prediction task.

(b) An example of the sentiment analysis task.

Figure 8: Examples of the action prediction and sentiment analysis task. Given the labeled movie clip, character profile and long context, the answers of different LLMs to the question prompt are also illustrated.

For the action prediction in Figure. 8(a), the answer " 吕受益接下来会做出摘下口罩的动作。(Lu Shouyi will then proceed to take off his mask.)" necessitates the model's ability to scrutinize the film and assimilate the visual cues pertaining to the mask for precise interpretation. Evidently, not only the conventional text-based LLMs but also GPT-4 and our MovieGPT fall short of accomplishing this task.

Regarding the task of sentiment analysis in Figure. 8(b), given the finite number of sentiment categories, this task is evaluated using the accuracy metric. It is observable that in the absence of multimodal information, the purely text-based LLMs, Bard and Claude2, still fail to provide accurate responses; whereas, the remaining three models all deliver correct answers.

Moreover, as shown in Figure 9, even with comprehension of the video content, it remains exceedingly challenging for a human to address the question " 这一举动也跟窗帘旁白挂着的星星之火可以燎原呼应了。(This action also echoes with the small spark hanging by the curtains, suggesting a potential wildfire, as mentioned in the stage directions.)" This necessitates a profound understanding of the film, explaining why all LLMs uniformly responded such as "the pressure of reality, and the unreality of the dream", which underscores the complexity of deriving nuanced interpretations from multimedia content.

A.1.2 Selection of Movies and Main Characters

Our selection was guided by a goal to ensure diversity in genres and historical span. We also focused on movies with strong narratives, clearly defined main characters, and a substantial amount of dialogue, as these elements are crucial for multimodal characterQA. Starting with a broad pool of 200 movies across various genres, we employed a meticulous review process by three annotators to identify movies meeting these criteria. Movies with weak narratives or lacking 3-5 main characters were excluded. We further filtered out movies with less than 50 lines of dialogue among the main characters. This rigorous process ultimately resulted in a curated list of 101 movies for our CharacterQA dataset.

For each movie, we chose the top 10 ranked characters from each movie's cast list, ensuring they had significant dialogue interaction (at least 50 lines), since less conversations of other secondary characters are

Figure 9: An example of the logical analysis task. Given the labeled movie clip, character profile and long context, the answers of different LLMs to the question prompt are also illustrated.

not enough to support multimodal characterQA. We then refined this selection by focusing on characters who were central to the movie's main storyline, resulting in selecting 3-5 movie characters for each movie. This methodical approach helped us create a robust and relevant dataset that accurately represents main characters in each movie. Finally, 398 main characters are obtained.

A.1.3 The Alignment of Dialogues and its Timestamps.

To obtain accurately matched dialogue transcripts from the movies, we utilized Alibaba's 'Tongyi Tingwu' software for audio-to-text transcription. This software provides dialogue texts with corresponding timestamps. However, the accuracy of the transcribed content was not always perfect. To address this, we engaged annotators to review and correct the transcriptions against the actual movie dialogues. A second round of validation by another annotator ensured the high alignment accuracy of the final dialogue texts, which further ensures the reliability of our data. Whether it's dialects or standard Mandarin, the manually processed dialogue texts exhibit extremely high matching rates with the movie dialogues. It's worth noting that our model does not have audio input, so the impact of dialects on performance is not as significant.

A.1.4 Quality Control for Crowd Worker Annotations

All annotations are conducted on an enterprise crowdsourcing platform by proficient Chinese workers with a solid educational background, ensuring accurate comprehension and precise description of video content. It's important to note that workers must have completed at least 100 prior tasks on the platform with a minimum approval rate of 95%. Additionally, we perform daily spot checks on worker annotations to ensure relevance to the videos. For validation of modified script dialogues, we enlist three native Chinese speakers who verify accuracy by watching 101 Chinese movies, confirming dialogue correctness and alignment with timestamps and characters. Corrections are made for any inconsistencies. Character profiling validation involves randomly assigning 398 main movie characters to 50 individuals, each assessed by two people familiar with Chinese and relevant movies. Ratings are provided, with consensus satisfaction levels over 85% considered a pass; otherwise, reassessment by annotators is conducted. Validation for the five video QA tasks involves reassigning annotated QA pairs to two additional individuals who assess consistency with the movie plot and characters, making modifications for any inconsistencies.

A.2 Experiments

In this section, extensive experiments will be provided about our CharacterQA dataset, including the training details, comparison of different evaluation metrics, and supplementary ablation studies.

A.2.1 Training Details

Table 8: Training hyperparameters for character conversation fine-tuning stage and character QA instruction tuning stage.

Hyperparameters	**Conversation Fine-Tuning**	**QA Instruction Tuning**
GPU type	8 $\times$ A6000	8 $\times$ A6000
Max token length	1024	1,024
Batch size of text instruction data	-	128
Optimizer	AdamW	AdamW
Learning rate	2e-4	2e-5
Learning rate decay	Cosine	Cosine
Adam ϵ	1e-6	1e-6
Adam β	(0.9, 0.98)	(0.9, 0.999)
Epoch	2	5
Weight decay	0.001	0.0001

Our MovieGPT is trained in two stages: the character conversation fine-tuning stage and the character QA instruction tuning stage, during which we freeze the parameters of the visual module and Bloom-7B, fine-tuning the latter with LoRA (Hu et al., 2021). LoRA allows us to indirectly train the dense layers in neural networks by optimizing the rank-decomposition matrices of dense layers during the adaptation process, while keeping the pre-trained weights unchanged.

The character conversation fine-tuning stage aims to familiarize the model with the multimodal character conversation task. In this stage, the visual module receives frames from the 60 second movie clip, while the text input contains long context, character profiles, and the prompt words " 该角色在该场景下<角色名>接下来要说什么话? (What is <Character Name> going to say next in this scenario?)". The output of the visual module and the text inputs will be fed into Bloom-7B together to generate the prediction for the next dialogue of the corresponding character. The loss function that maximizes the likelihood estimation between the prediction and groundtruth dialogue is adopted. The detailed training parameters for this stage is demonstrated in Table. 8.

The character QA instruction tuning stage aims to enable the model's abilities of the specific task. Particularly, the labeled data of our five designed tasks (dialogue prediction, action prediction, relationship judgment, logical analysis, and sentiment analysis) is adopted to train our MovieGPT, i.e., frames from 60 second movie clips are fed into the frozen visual module, the specific task-related question is the text input. Given the output visual feature and the question, Bloom-7B produces the answer to the question, with the loss function again focusing on maximizing the likelihood estimation between the predicted response and the correct answer. The detailed training parameters for this stage is demonstrated in Table. 8.

The underlying principle of the model training of our MovieGPT is as follows:

1. The first training stage is designed to enable the model to learn personalized representations of characters within the movie, which allows the model to grasp the intricacies of the plot and facilitate accurate dialogue predictions.

2. In the second stage, the model can actively engage in question-answering tasks tailored to various personalized scenarios, leveraging its understanding gained from analyzing movie clips based on provided instructions.

A.2.2 Evaluation Details

To comprehensively verify the character conversation performance, apart from the evaluation metrics, we also adopt human annotators to evaluate the open-domain VideoQA tasks in our CharacterQA dataset. For human evaluation of the responses of different models, we adopted a common method, engaging several expert annotators for evaluation to maintain opinion alignment. Concretely, groups of five native Chinese speakers independently reviewed the relevant movie and assessed answer consistency. In cases of disagreement, group discussion started, and three additional evaluators will assess answer consistency. This iterative process continued until unanimous agreement was reached within the group, thereby maximizing the precision of our consistency ratings.

A.2.3 Comparison of Different Evaluation Metrics

CIDEr Score	DP	AP	RJ	SA	LA
Bard	0.79	11.49	39.57	-	15.43
Claude-2	2.46	23.13	66.34	-	38.15
GPT-3.5	2.61	16.76	83.23	-	44.13
GPT-4	6.75	33.51	**91.22**	-	**51.96**
MovieGPT	**21.89**	**41.02**	78.30	-	19.51

Table 9: The CIDEr score of different models on five tasks. “DP” stands for dialogue prediction, “AP” stands for action prediction, “RJ” stands for relationship judgment, “SA” stands for sentiment analysis, “LA” stands for logical analysis.

ROUGE-L Score	DP	AP	RJ	SA	LA
Bard	0.19	0.00	12.34	-	0.00
Claude-2	0.57	0.04	28.09	-	0.01
GPT-3.5	0.53	0.17	37.16	-	0.01
GPT-4	1.84	0.33	**48.26**	-	0.02
MovieGPT	**7.67**	**0.75**	33.89	-	0.00

Table 10: The ROUGE-L score of different models on five tasks.

In the paper, we have furnished a comprehensive comparison of BLEU scores across various tasks. Given the substantial challenges inherent in evaluating open-domain responses, where multiple correct answers are possible, especially for our multimodal character QA, we also further present the CIDEr, ROGUE-L and human evaluation results of different models across five tasks in Table. 9, Table. 10 and Table. 11, respectively. The inputs for the text-only LLMs during the experiments include the long context preceding the movie clips, the text dialogue of the movie clips, character profiles, and the question of the corresponding task for the current clip.

Note that the BLEU, CIDEr, and ROUGE-L metrics measures the qualities of answers from different perspective. For example, ROUGE-L calculates the longest common subsequence between the answer (C) and the groundtruth sentence (S), as shown in Equ. 1:

$$\begin{aligned} R_{LCS} &= \frac{LCS(C,S)}{\text{len}(S)} \\ P_{LCS} &= \frac{LCS(C,S)}{\text{len}(C)} \\ ROUGE-L = F_{LCS} &= \frac{(1+\beta^2)\, R_{LCS}P_{LCS}}{R_{LCS}+\beta^2 P_{LCS}} \end{aligned} \tag{1}$$

where R_{LCS} represents recall, while P_{LCS} represents precision, and F_{LCS} is ROUGE-L. Typically, β is set to a large number, so F_{LCS} almost only considers R_{LCS} (i.e., recall). Note that when β is large, F_{LCS}

will focus more on R_{LCS} than P_{LCS}. If β is very large, the P_{LCS} term can be disregarded. Obviously, this metrics is very strict for open-domain QA since it requires responses and answers to be identical in the longest possible sequence. Most tasks begin with a common sequence for ROUGE-L scores, e.g., "The character will next" for action prediction, "The character will next say" for dialogue prediction, and "The relationship between character 1 and character a is" for relationship prediction; while logical analysis scored lower due to the lack of a common sequence.

As shown in Table. 9 and Table. 10, the CIDEr and ROUGE-L scores were in good agreement with the BLEU scores in terms of overall trends, although there were some differences in the spread between the different model effects. Furthermore, as previously noted, the ROUGE-L scores for all models approached zero, attributable to the inherent challenges of this open-domain task and the constraints of the metric itself.

For the more accurate human evaluations shown in Table. 11, different models obtaining A or B scores also showed the same results. This suggests that metrics such as BLEU-4 and CIDEr, despite their limitations, are still valuable in assessing answer quality. Furthermore, our MovieGPT still achieves the best results over all tasks, further confirming its effectiveness.

Model	Dialogue Prediction				Action Prediction				Relationship Judgment				Logical Analysis			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
Bard	0	3	14	483	11	74	41	374	137	137	129	97	11	43	55	391
Claude2	2	7	32	459	38	109	72	281	220	121	122	37	21	62	78	339
GPT-3.5	4	8	40	448	21	102	74	303	289	111	81	19	37	83	72	308
GPT-4	9	11	51	429	63	146	53	238	317	109	63	11	59	71	86	284
MovieGPT	68	54	163	215	82	90	157	171	254	86	76	84	17	81	45	357

Table 11: The human evaluation results of different models on four open-domain tasks, where ratings are assigned across four levels: very consistent (A), somewhat consistent (B), fairly consistent (C), and not consistent (D).

A.2.4 Supplementary Ablation Studies

Movie Clips	Role Profile	Long Context	DP	AP	RJ	SA	LA
✓			3.12	7.88	46.27	20.06%	1.97
	✓		0.71	5.18	29.36	13.21%	1.22
		✓	9.87	14.21	60.43	27.64%	2.01
	✓	✓	13.13	26.43	70.44	31.18%	3.32
✓		✓	11.46	28.13	66.21	36.91%	2.05
✓	✓		15.28	29.21	73.62	37.29%	4.45
✓	✓	✓	15.28	29.21	73.62	37.29%	4.45

Table 12: Our method in comparison with others through ablation experiments under different scenarios.

In this section, we have conducted extensive ablation studies on our CharacterQA dataset. For all results in this section, the BLEU score is adopted for dialogue prediction, action prediction, relationship judgment, and logical analysis, and the Accuracy metric is adopted for sentiment analysis.

Models	Training Setting	DP	AP	RJ	SA	LA
NExT-GPT	In-context Learning	0.29	5.34	21.05	11.63%	1.46
Video-LLaVA	In-context Learning	0.45	6.69	46.38	26.86%	1.87
NExT-GPT	Two-stage Training	4.67	11.64	57.98	23.64%	2.31
Video-LLaVA	Two-stage Training	6.44	17.63	61.04	31.86%	3.37
MovieGPT	Two-stage Training	15.28	29.21	73.62	37.29%	4.45

Table 13: The comparison between our MovieGPT and other multimodal LLMs on five tasks.

Moreover, to underscore the complexity of the CharacterQA dataset and affirm the efficacy of MovieGPT, we embarked on comparative experiments with other well-regarded multimodal LLMs. Our selection was constrained by the scarce availability of open-source multimodal LLMs capable of processing both Chinese language and video inputs. For instance, in our preliminary evaluations, Video-LLaMA (Zhang et al., 2023) exhibited proficiency in handling video temporal information but fell short in accommodating Chinese conversational contexts. Consequently, NExT-GPT (Wu et al., 2023) and Video-LLaVA (Lin et al., 2023), two popular open-source multimodal platforms, were chosen for evaluation.

Specifically, the results of both NExT-GPT and Video-LLaVA, when performing in-context learning only and when performing the identical two-stage training as our MovieGPT, are shown in the Table. 13. As we can see, with only in-context learning, Video-LLaVA and NExT-GPT exhibit suboptimal performance across all tasks, even underperforming some text-only LLMs. This can be attributed to their inadequate comprehension of video content, which may exacerbate errors in the absence of training tailored to multimodal character conversation. After two stages of training, both NExT-GPT and Video-LLaVA show significant improvement over their results with in-context learning only. However, their results are still inferior to our MovieGPT over all tasks.

Length of movie clip	DP	AP	RJ	SA	LA
10s	14.68	28.12	73.06	35.22%	3.48
60s	15.28	29.21	73.62	37.29%	4.45
300s	16.97	30.58	76.29	38.62%	6.67

Table 14: The performance of our MovieGPT with the movie clip of different lengths.

Furthermore, we attempt ablation experiments with different video lengths in Table. 14, which shows marked improvement when the duration of the movie clip increases to 300 seconds, confirming the necessity of incorporating "visual context" is crucial for multimodal characterQA. However, much longer videos will face a dilemma of high frame extraction computational costs, and less frame extraction will result in more temporal information loss. The visual processing capabilities of existing multimodal LLMs for long videos are also limited (Zhang et al., 2023; Liu et al., 2023b) (For the common multimodal LLMs like Video-LLaMA (Zhang et al., 2023) and mPLUG-owl (Ye et al., 2023c), the lengths of video input are usually less than 3 minutes due to their inferior visual capacities.). Designing a model capable of rapidly processing longer movie clips to enhance multimodal character QA performance is a future research direction for us.