FriendsQA: A New Large-Scale Deep Video Understanding Dataset with Fine-grained Topic Categorization for Story Videos

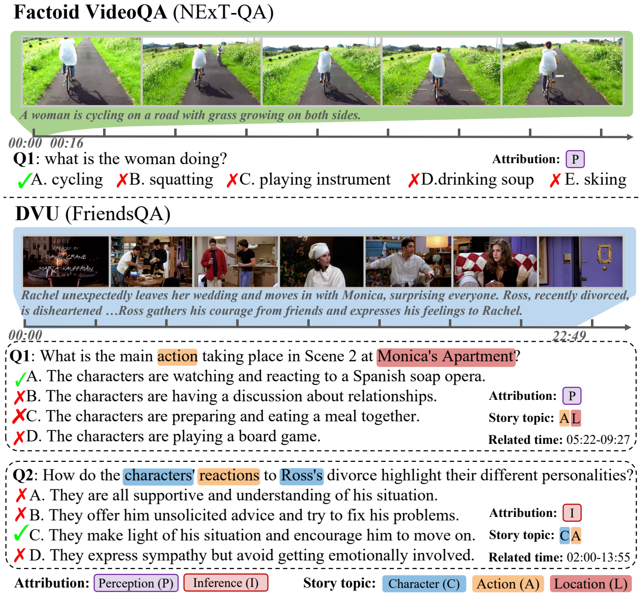
Zhengqian Wu[1,3,4*], **Ruizhe Li**[1,3,4*], **Zijun Xu**[2,3,4],
Zhongyuan Wang[1,2,3,4], **Chunxia Xiao**[1,2,3,4], **Chao Liang**[1,2,3,4†]

[1]School of Computer Science, Wuhan University, China
[2]School of Cyber Science and Engineering, Wuhan University, China
[3]National Engineering Research Center for Multimedia Software, Wuhan University, China
[4]Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China
{2023202110068, 2020300004016, 2019300003083, cxxiao, cliang}@whu.edu.cn, wzy_hope@163.com

Abstract

Video question answering (VideoQA) aims to answer natural language questions according to the given videos. Although existing models perform well in the factoid VideoQA task, they still face challenges in deep video understanding (DVU) task, which focuses on story videos. Compared to factoid videos, the most significant feature of story videos is storylines, which are composed of complex interactions and long-range evolvement of core story topics including characters, actions and locations. Understanding these topics requires models to possess DVU capability. However, existing DVU datasets rarely organize questions according to these story topics, making them difficult to comprehensively assess VideoQA models' DVU capability of complex storylines. Additionally, the question quantity and video length of these dataset are limited by high labor costs of handcrafted dataset building method. In this paper, we devise a large language model based multi-agent collaboration framework, StoryMind, to automatically generate a new large-scale DVU dataset. The dataset, FriendsQA, derived from the renowned sitcom *Friends* with an average episode length of 1,358 seconds, contains 44.6K questions evenly distributed across 14 fine-grained topics. Finally, We conduct comprehensive experiments on 10 state-of-the-art VideoQA models using the FriendsQA dataset.

Code — https://github.com/nercms-mmap/FriendsQA
Extended version — https://arxiv.org/abs/2412.17022


Factoid VideoQA (NExT-QA)
A woman is cycling on a road with grass growing on both sides.
00:00
00:16
Attribution: P
DVU (FriendsQA)
Rachel unexpectedly leaves her wedding and moves in with Monica, surprising everyone. Ross, recently divorced, is disheartened …Ross gathers his courage from friends and expresses his feelings to Rachel.
00:00
22:49
Attribution: P
Story topic: A L
Related time: 05:22-09:27
Attribution: I
Story topic: C A
Related time: 02:00-13:55
Attribution: Perception (P) Inference (I)
Story topic: Character (C) Action (A) Location (L)

Q1: what is the woman doing?
✓A. cycling ✗B. squatting ✗C. playing instrument ✗D.drinking soup ✗ E. skiing

Q1: What is the main action taking place in Scene 2 at Monica's Apartment?
✓A. The characters are watching and reacting to a Spanish soap opera.
✗B. The characters are having a discussion about relationships.
✗C. The characters are preparing and eating a meal together.
✗D. The characters are playing a board game.

Q2: How do the characters' reactions to Ross's divorce highlight their different personalities?
✗ A. They are all supportive and understanding of his situation.
✗ B. They offer him unsolicited advice and try to fix his problems.
✓C. They make light of his situation and encourage him to move on.
✗ D. They express sympathy but avoid getting emotionally involved.

Figure 1: Comparisons of factoid VideoQA and DVU.

Introduction

Video question answering (VideoQA) aims to answer natural language questions based on given videos, supporting advanced applications like video grounding (Ren et al. 2024) and video chatbots (Jin et al. 2024). Early researches on factoid VideoQA focus on questions about actions or objects in videos (Yu et al. 2019). With the popularity of video language model (VLM) and multimodal large language model (MLLM) technologies, significant progress has been made in recent years (Yu et al. 2023).

However, VideoQA models' performance significantly declines on deep video understanding (DVU) task. For instance, VideoChat2 achieves 61.70% accuracy on the factoid NExT-QA dataset but drops to 44.05% on our DVU FriendsQA dataset. As illustrated in Figure 1, the key reasons include: First, regarding question attributes, in addition to perception questions that inquire about visual cues, DVU also includes inference questions that assess the understanding of storylines, offering more diverse questions. Second, for the same perception questions, factoid VideoQA involves short-range shot-level perception in factoid video depicting simple events without complex narrative interactions, whereas DVU involves long-range scene-level (Liang et al. 2009; Xu, Wei, and Wu 2023) perception in story videos encompass

[*]These authors contributed equally.
[†]Chao Liang is the corresponding author.

Dataset	Venue	Fine-grained topic distribution			Question Scale			Difficulty Measure	Cross Episodes
		# Fin. Top.	Gin.	Ent.	# Que.	Vid. Len. (s)	# Que.×Vid. Len. (Ks)		
MovieQA	CVPR'16	6	0.819	2.713	14.9K	202.7	3,020.2	✗	✗
TVQA	EMNLP'18	8	0.821	2.873	144.9K	76.2	11,041.4	✗	✗
TVQA+	ACL'20	5	0.789	2.660	29.4K	61.5	1,808.1	✗	✗
HLVU (DVU 22&23)	ICMR'20	6	0.773	2.548	455	106 / 4,907	1,010.5	✗	✗
DramaQA	AAAI'21	-	-	-	17.9K	3.6 / 91.8	429.9	✓	✗
DeepMaven	EACL'23	-	-	-	1K	3,102	3,102.0	✗	✗
CinePile	CVPRW'24	-	-	-	**200K**	160	32,000.0	✗	✗
MovieChat-1K	CVPR'24	4	0.701	2.203	19.0K	564	10,716.0	✗	✗
FriendsQA (ours)		**14**	**0.927**	**3.794**	44.6K	**1,358 / 5,390**	**98,874.8**	✓	✓

Table 1: Comparisons of existing DVU datasets. Fine-grained topic distribution consider the number of fine-grained topics exceeding 5% of the dataset (# Fin. Top.) and the balance degree of topic distribution. The Gini index (Gin.) and entropy (Ent.) are employed to measure the distribution's balance. Question scale is compared by analyzing the number of questions (# Que.), the average video length (Vid. Len.), and their product. The proposed FriendsQA dataset offers additional features, including difficulty measure and cross episodes questioning. The figures around the "/" represent two distinct video input lengths.

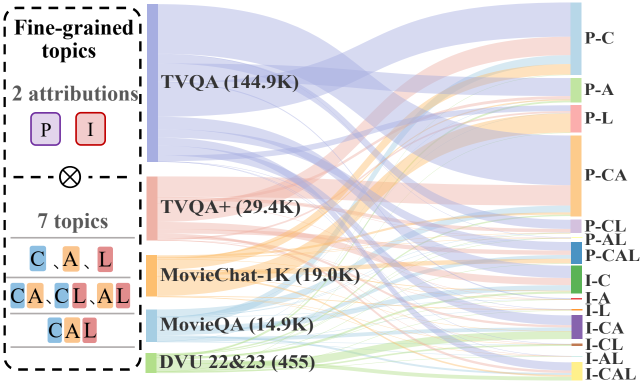


Figure 2: The distribution of 5 datasets questions across the 14 fine-grained topics, based on 7 topics (character (C), action (A), location (L) and their combinations) and 2 cognitive attributions, i.e., perception (P) and inference (I).

intricate storylines. In addition, characters and locations in DVU have specific identities (Sang et al. 2011), i.e., Ross and Monica's apartment. Due to these factors, the DVU task is more challenging. The core of DVU task lies in understanding storylines with long-range evolvement, composed of story topics (Guo, Liang, and Wang 2023), i.e. characters, actions, locations and their combinations.

Although story topics are essential to analyzing storylines, the most of existing DVU datasets (Lei et al. 2018; Tapaswi et al. 2016; Song et al. 2024) do not organize questions based on them. We use Gemini 1.5 Pro[1] to categorize questions from 5 classic DVU datasets (Please refer to Appendix A for more details.) into 14 fine-grained topics we refer to as in this paper, i.e., 7 story topics, character, action, location and their combinations attached to 2 question attributions, perception and inference (Ammanabrolu et al. 2021). The results are shown in Figure 2. We find that these datasets tend to concentrate on a limited type of fine-grained topics. This situation make it difficult to provide a comprehensive evaluation of VideoQA models' DVU capability.

Furthermore, complex storylines involve long-range evolvement, necessitating both numerous questions to cover the entire storyline and sufficiently video length to convey the storyline clearly. However, as shown in Table 1, existing DVU datasets either pair a large number of questions with short videos, such as TVQA (Lei et al. 2018), or feature longer videos with fewer questions, such as DeepMaven (Fung et al. 2023). This limitation prevents them from evaluating the model's understanding of storylines in both breadth and depth simultaneously.

To address the above two challenges, we propose StoryMind, a large language model (LLM) based multi-agent collaboration framework, comprising a generator to generate questions with 14 fine-grained topics and two reviewers to remove low-quality ones. We provide the generator with detailed topic explanations and manually constructed examples to guide question generation. To ensure balanced topic coverage, generator will iteratively generate questions until the number of questions for each fine-grained topic reaches the predetermined threshold. To ensure dataset quality, only the question and its answer unanimously deemed reasonable and accurate by both reviewers will be retained. Moreover, as illustrated in Table 1, we perform difficulty measure for each question to accurately evaluate the DVU capability of various VideoQA models, and introduce cross-episode questions to augment the challenge of the dataset by covering more complex storyline.

We apply StoryMind to the popular sitcom *Friends*, comprising 234 episodes with average video length of 1,358s across ten seasons. This results in a new large-scale DVU dataset, FriendsQA, with the most diverse and balanced fine-grained topics. As studied in Table 1, this dataset includes 44.6K questions evenly distributed across 14 fine-grained topics. Thereafter, we comprehensively evaluate the DVU capbility of 10 state-of-the-art (SOTA) models on FriendsQA, encompassing both VLM-based methods and newly proposed MLLM-based methods.

[1] https://aistudio.google.com/

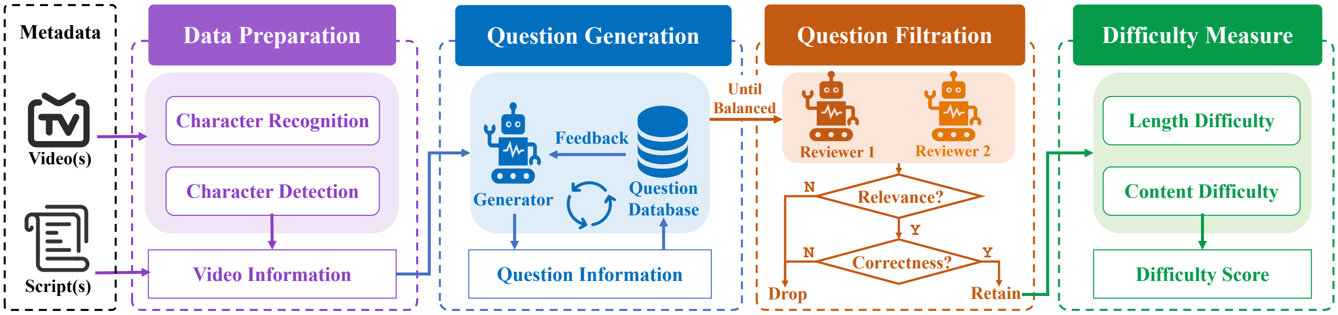


Figure 3: The workflow diagram of the multi-agent collaboration framework StoryMind.

In summary, our contributions are three-fold:

- We devise StoryMind, a LLM-based multi-agent collaboration framework to automatically generate and review large-scale questions with diverse fine-grained topics.
- We construct FriendsQA, a large-scale DVU dataset comprising 44.6K questions with 14 fine-grained topics tailored for thorough DVU evaluation
- We conduct extensive evaluations of 10 SOTA VideoQA models on FriendsQA.

Related Work

In this section, we briefly overview the VideoQA datasets, encompassing factoid VideoQA and DVU datasets.

Factoid VideoQA Datasets

Factoid VideoQA dataset (Xu et al. 2017; Yu, Kim, and Kim 2018; Garcia et al. 2020; Li et al. 2020; Yang et al. 2021; Mangalam, Akshulakov, and Malik 2023) mainly focus on simple visual fact in short-range, such as object recognition, action recognition, spatial and temporal understanding in shot level. ActivityNet-QA (Yu et al. 2019) focuses on actions recognition, spatial relationships, and temporal relationships. It contains 58K human-annotated QA pairs on 5.8K videos from ActivityNet (Fabian Caba Heilbron and Niebles 2015). NExT-QA (Xiao et al. 2021) delves videos featuring object interaction, and there are 52K manually annotated questions including temporal, and descriptive questions. A notable distinction of these datasets is that they focus on short-range understanding of video with out complex storyline. This is a major factor contributing to the significant performance gap between factoid VideoQA and DVU.

DVU Datasets

Story videos, such as TV shows, are composed of complex interactions and long-range evolvement of core story topics. Understanding these topics requires models to possess deep video understanding (DVU) capability. Early work, such as PororoQA (Kim et al. 2017), focused on scene-dialog storytelling without complex storylines. Recent research has increasingly shifted towards understanding complex storylines in story video. However, as illustrated in Figure 2, the majority of current DVU datasets (Tapaswi et al. 2016; Lei et al. 2018, 2020; Song et al. 2024; Curtis et al. 2020) do not organize questions based on story topics. This phenomenon indicates existing datasets are difficult to provide a comprehensive evaluation of the VideoQA models' capability of DVU.

In addition, As shown in Table 1, some DVU datasets contain a large number of questions but are paired with short videos less than 5 minute, e.g., TVQA (Lei et al. 2018). Conversely, the other datasets encompass long videos (approximately 90 minutes). e.g. HLVU (Curtis et al. 2020). Recently, the DVU 2022 (Curtis et al. 2022) and DVU 2023 (Curtis et al. 2023) grand challenges[2] have comprehensively addressed on HLVU dataset. Nevertheless, These datasets often have a limited number of questions. For example, HLVU and DeepMaven (Fung et al. 2023) have fewer than 1K questions, and the largest, MovieChat-1K (Song et al. 2024), has only 19K questions. It hinders the evaluation of the model's understanding of storylines in terms of both breadth and depth simultaneously.

Particularly, There are some movie/TV-based datasets (Yang and Choi 2019; Ma, Jurczyk, and Choi 2018; Chen et al. 2021) for story understanding in the NLP domain. Specifically, compared to script-based QA dataset also named FriendsQA (Yang and Choi 2019), ours has two key differences: (1) In the question generation phrase, we explicitly add spatio-temporal information, such as timeline and character detection boxes obtained from video, to better suit the DVU task. (2) In the question answering phase, we provide video with QA pair as the basis of answering questions.

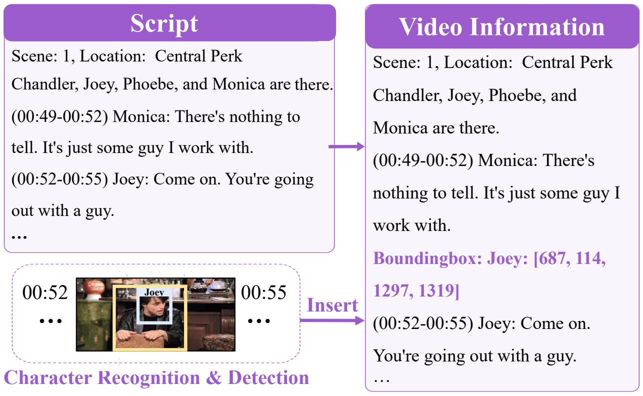
StoryMind

To reduce the labor costs, we propose StoryMind, a multi-agent collaboration framework. It can automatically generate large-scale questions with conprehensive and balanced fine-grained topics. As shown in Figure 3, it mainly consists of four stages, data preparation, question generation, question filtration and difficulty measure.

Data Preparation

The data preparation phase is responsible for converting the metadata into video information.

[2] https://sites.google.com/view/dvuchallenge2023/home



Scene: 1, Location: Central Perk
Chandler, Joey, Phoebe, and Monica are there.
(00:49-00:52) Monica: There's nothing to tell. It's just some guy I work with.
(00:52-00:55) Joey: Come on. You're going out with a guy.
…

Scene: 1, Location: Central Perk
Chandler, Joey, Phoebe, and Monica are there.
(00:49-00:52) Monica: There's nothing to tell. It's just some guy I work with.
Boundingbox: Joey: [687, 114, 1297, 1319]
(00:52-00:55) Joey: Come on. You're going out with a guy.
…

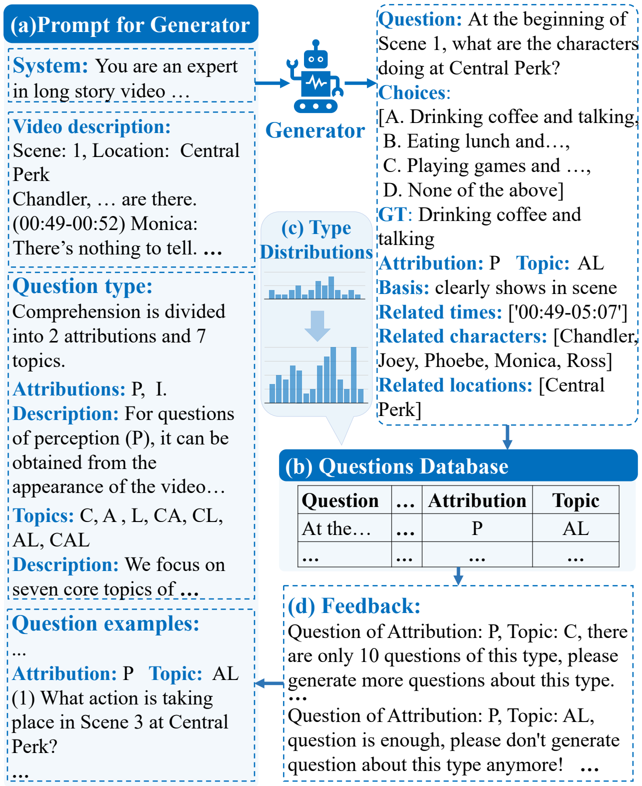
Figure 4: The flowchart of data preparation.

Metadata The script and episode video represent the textual structure and visual presentation of the storyline, respectively. Therefore, we use them as the metadata for question generation. The script from PAINS dataset (Niu et al. 2023) contains multiple scenes, with each scene detailing the location, the main characters involved, and aligning the dialogue with the video timeline, as shown in Figure 4. It provides the generator with a complete storyline evolvement ensuring that the generated questions are more accurate and closely aligned with the storyline. The timeline allows the generator to accurately generate the start and end timestamps corresponding to each question, facilitating the subsequent calculation of difficulty scores.

Video Information We apply the shot-based instance search method (Li et al. 2023) to detect and identify characters in the episode with, obtaining detection bounding boxes. This additional visual information allows the generator to generate questions with positional details, i.e., left, and right, thereby increasing questions complexity. Finally, we insert the bounding boxes into the script according to the timeline, as shown in Figure 4, creating a crucial component of the question generation prompts, which we refer to as the *video information*. Notebly, to enquire the more complicated storylines, we specifically designed cross-episode questions by concatenating 4 consecutive episodes[3].

Question Generation

The question generation stage aims to produce questions with diverse and balanced fine-grained topics. We introduce Gemini 1.5 Pro as generator and prompt the generator with *video information*, *descriptions of the fine-grained topics* and *questions examples*, as illustrated in Figure 5(a). Video information is the output of data preparation stage that record the script, timeline, and character positions. Descriptions of the fine-grained topic contains the description of 2 attributions (P, I) and 7 topics (C, A, L, CA, CL, AL and CAL). Questions examples include manually designed examples for each fine-grained topic to assist the generator in better understanding the fine-grained topics. All of them help the generator understand the categorization of fine-grained topics and generate corresponding questions.



System: You are an expert in long story video …

Video description: Scene: 1, Location: Central Perk
Chandler, … are there.
(00:49-00:52) Monica: There's nothing to tell. …

Question type: Comprehension is divided into 2 attributions and 7 topics.
Attributions: P, I.
Description: For questions of perception (P), it can be obtained from the appearance of the video…
Topics: C, A , L, CA, CL, AL, CAL
Description: We focus on seven core topics of …

Question examples:
…
Attribution: P Topic: AL
(1) What action is taking place in Scene 3 at Central Perk?
…

Question: At the beginning of Scene 1, what are the characters doing at Central Perk?
Choices:
[A. Drinking coffee and talking,
B. Eating lunch and…,
C. Playing games and …,
D. None of the above]
GT: Drinking coffee and talking
Attribution: P Topic: AL
Basis: clearly shows in scene
Related times: ['00:49-05:07']
Related characters: [Chandler, Joey, Phoebe, Monica, Ross]
Related locations: [Central Perk]

Question	…	Attribution	Topic
At the…	…	P	AL
…	…	…	…

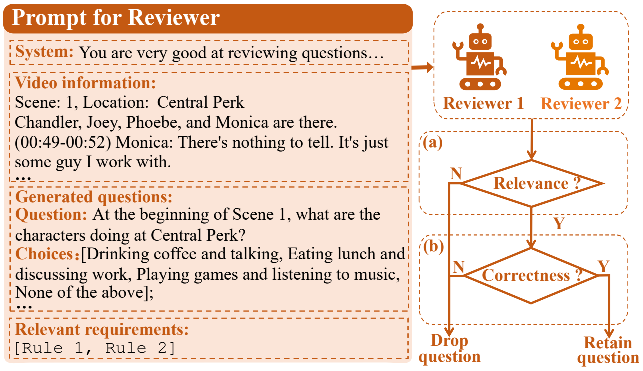
Question of Attribution: P, Topic: C, there are only 10 questions of this type, please generate more questions about this type.
…
Question of Attribution: P, Topic: AL, question is enough, please don't generate question about this type anymore! …

Figure 5: The iterative flowchart of question generation.

In addition, to ensure a balanced distribution of questions across each fine-grained topic, we require the generator to specify the fine-grained topic categorization of each question and save the information into question database (Figure 5(b)). This allows database to track the distribution across all fine-grained topics (Figure 5(c)). To prevent the generator from focusing solely on one topic, we devise a feedback mechanism. The database checks the number of questions for each fine-grained topic against a same threshold. If a topic falls below the threshold, the database will provide feedback to generator with generating more questions for that topic. Otherwise, the opposite feedback will be given (Figure 5(d)). The generator will iteratively generate questions until all topics reach the threshold.

As illustrated in generated result of Figure 5, we require the generator to output additional information for each question, such as choices list and ground truth (GT). Since the video information contains a complete timeline, character and location information, we require the generator to specify the related time, characters, and locations associated with each question. These will be used for subsequent difficulty measure. Please refer to Appendix B.1 for more details.

[3] We choose 4 episodes because the scripts from 4 episodes reach the generator's context limit.



Prompt for Reviewer

System: You are very good at reviewing questions…

Video information:
Scene: 1, Location: Central Perk
Chandler, Joey, Phoebe, and Monica are there.
(00:49-00:52) Monica: There's nothing to tell. It's just some guy I work with.
…

Generated questions:
Question: At the beginning of Scene 1, what are the characters doing at Central Perk?
Choices:[Drinking coffee and talking, Eating lunch and discussing work, Playing games and listening to music, None of the above];
…

Relevant requirements:
`[Rule 1, Rule 2]`

Figure 6: The flowchart of question filtration.

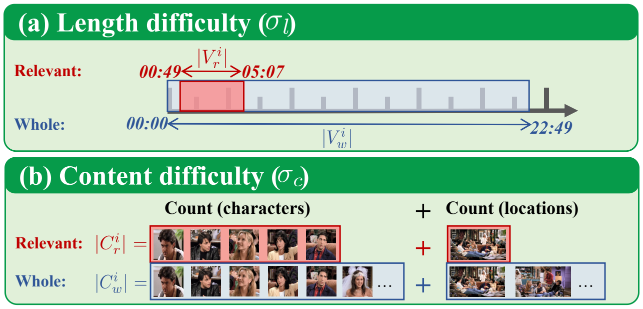


Figure 7: Illustration of two difficulty factors.

Question Filtration

Question filtration aims to filter out incorrect questions. We introduce Gemini 1.5 Pro and Claude 3.5 Sonnet[4] as two reviewers to automate the process. The prompt is meticulously designed for them as shown in Figure 6 (For complete prompt, please refer to Appendix B.2), which mainly consists of three parts, *video information*, *generated questions* and *relevant requirements*. Video information is the output of data preparation. It ensures reviews to be on the same level of obtaining storylines as the generator, allowing them to fairly assess the questions' accuracy. Generated questions is the output of question generation phase. Reviewers evaluate the generated questions based on the video information. Relevant requirements include two rules:

- **Rule 1** The question must be relevant to the video and can be answered with GT from the generator.
- **Rule 2** Among the 4 choices, there must be only one correct answer, and three wrong answers.

If the both rules are met, the output for the correctness of the question is `True`; otherwise, it is `False` (Figure 6(a)). Additionally, The unique correct answer selected independently by each of the two reviewers must be identical and consistent with GT (Figure 6(b)).

Difficulty Measure

Difficulty measure aims to assign a difficulty score to each question to better assess the DVU capabilities of VideoQA models (Zhang et al. 2022). We propose 2 difficulty factors, i.e., video length and content. Let $|V_w^i|$ and $|V_r^i|$ denote the length of the whole video and relevant video for i-th question. $|C_w^i|$ and $|C_r^i|$ denote the instance (including characters and locations obtained from video information) number of the whole video and relevant video to represent content. Therefore, we define 2 difficulty scores:

Dataset	Manul	Revision	Single		Cross	
			# Num	Ratio	# Num	Ratio
FriendsQA-S1	✗	✗	3,795	100%	995	100%
FriendsQA-M w/o R.	✓	✗	3,475	91.57%	894	89.85%
FriendsQA-M	✓	✓	3,584	94.44%	905	90.95%

Table 2: Comparison of datasets on first season in different filtration and type. w/o R. indicates revison are not included.

Type	Dataset		Difference
	FriendsQA-S1	FriendsQA-M	
single	33.45	33.60	0.15
cross	34.77	35.02	0.25

Table 3: The average accuracy and difference (%) of 10 SOTA models across single and cross-episode questions on FriendsQA-M and FriendsQA-S1.

- Length difficulty ($\sigma_l^i = |V_w^i|/|V_r^i|$). It represent the relationship of length between whole video and relevant video as shown in Figure 7(a). Larger value indicates smaller proportion of relevant video length, making it harder for models to capture the necessary information.
- Content difficulty ($\sigma_c^i = |C_w^i|/|C_r^i|$). It characterizes the relationship of video content between whole video and relevant video as shown in Figure 7(b). Larger value indicates less video content of relevant video, making it harder for models to capture the necessary instance.

Based on the two factors mentioned above, we use the following formula to calculate the overall difficulty score:

$$\sigma^i = \frac{\sigma_l^i}{\mu_l} + \frac{\sigma_c^i}{\mu_c} \qquad (1)$$

where μ_l and μ_c represent the average length and content scores for all questions, which is used to mitigate factor discrepancy. In the "Evaluation" section, we validate the effectiveness of our difficulty measure through experiments.

FriendsQA Dataset

Dataset Quality

Manual Verification Before presenting the dataset statistics, we assess the quality of the FriendsQA generated by the StoryMind. This evaluation ensures the validity of subsequent evaluation. Therefore, we first select 4,790 questions from the first season (denoted as FriendsQA-S1) and apply the same requirements as in the previous "Question Filtration" and manually revise questions with incorrect options

[4] https://claude.ai/

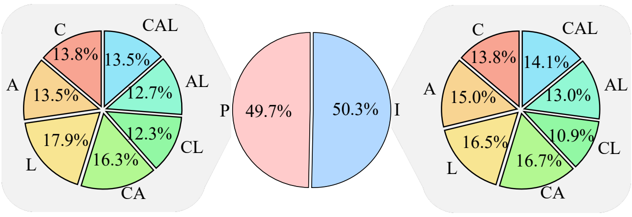


Figure 8: The distribution of fine-grained topics.

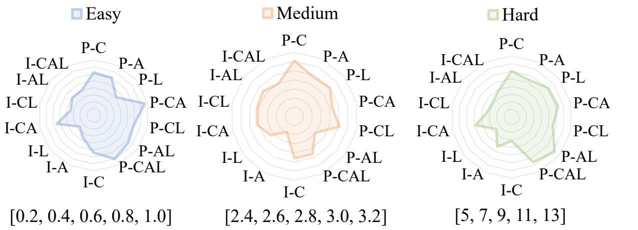
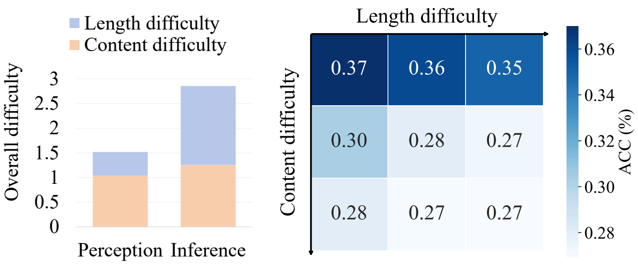


Figure 9: Overall difficulty score on different levels. The numbers in [·] are coordinate values for each circle.



(a) Average content and length difficulty for P and I questions.

(b) The impact of the 2 difficulties on the average accuracy of 10 models.

Figure 10: The influence of the two difficulties on accuracy

(e.g., the correct answer is A but B is selected). In the subsequent experiments, we use the manually verified dataset with revision (denoted as FriendsQA-M).

Manual v.s. Automatic The comparisons between FriendsQA-M and FriendsQA-S1 are shown in Table 2. For single-episode questions, 91.57% are retained directly without revision (denoted as FriendsQA w/o R.), and 94.44% are retained with revision. For cross-episode questions, these figures are 89.85% and 90.95%, respectively. Furthermore, we conduct analysis of the accuracy difference of 10 SOTA models (will be elaborated in the latter "Evaluation" section) on FriendsQA-S1 and FirendsQA-M, as shown in Table 3. The average differences between them are 0.15% for single-episode questions and 0.25% for cross-episode questions. These results demonstrate the high-quality of the FriendsQA and the effectiveness of the StoryMind.

Dataset Statistics

Dataset Scale We applied StoryMind to the classic sitcom *Friends*, which contains 234 episodes with average length of 1358s. It leads to a large-scale dataset, FriendsQA, with over 44.6K questions, covering 14 fine-grained topics. The number of single-episode and cross-episode questions are 35,222 and 9,470, respectively, in approximately a 4:1 ratio. Examples for 14 fine-grained topic in FriendsQA are elaborated in Appendix C.

Fine-grained Topic As illustrated in Figure 8, The proportion of P and I questions is nearly 50%, and for specific topics, like the CL with the lowest proportion of I questions, it still accounts for 10.9%. This indicates that the questions are evenly distributed across the 14 fine-grained topics. The distribution is the same for single and cross-episode question, which is detailed in Appendix D.

Difficulty Finally, to facilitate the analysis of results in conjunction with question difficulty, we sort the questions in ascending order of difficulty score σ^i and divide them into three levels, easy, medium and hard based on a 9:3:1 ratio. The average difficulty scores of the 14 fine-grained topics across different difficulty levels are shown in the Figure 9. It can be observed that the difficulty of P questions tends to be higher than that of I questions across all three levels.

Evaluation

Setting

Baseline We benchmark 10 SOTA models over the past two years and categorize these models into two groups:

- **MLLM** models including Chat-Univi (Jin et al. 2024), MA-LMM (He et al. 2024), MovieChat (Song et al. 2024), SeViLA (Yu et al. 2023), TimeChat (Ren et al. 2024), Video-ChatGPT (Maaz et al. 2024), VideoChat2 (Li et al. 2024), VideoLLaMA2(Cheng et al. 2024). Except for SeViLA, which uses FlanT5-XL 3B (Chung et al. 2022) as backbone, all other MLLM models, use various versions of the LLaMA model (Touvron et al. 2023a,b) as backbone. For these models, we consistently use the 7B parameter version of LLaMA as the backbone.
- **VLM** models including Vid-TLDR (Choi et al. 2024) and VIOLETv2 (Fu et al. 2023). We use the weights fine-tuned on MSRVTT-QA, the most commonly used VideoQA dataset, to ensure fairness.

Implementation Details We set models in zero-shot question-answering settings on FriendsQA, with official default configurations. Evaluation is performed with the maximum frames supported by each model while staying within 48GB capacity of RTX A6000 GPU. More implementation details of 10 tested models are elaborated in Appendix E.

Evaluation Metrics We use accuracy (Li et al. 2023) as metrics, which is calculated by dividing the number of correct answered questions by the total number of questions.

Evaluation Result

MLLM v.s. VLM Table 4 shows an overall result of 10 models. By comparing MLLM models and VLM models, we find most MLLM models achieve better performance. This may be due to the fact that MLLMs utilize LLM as backbone, having greater language comprehension abil-

Att.	Top.	MLLM								VLM		AVG
		Chat-UniVi	MA-LMM	MovieChat	SeViLA	TimeChat	Video-ChatGPT	VideoChat2	VideoLLaMA2	VIOLETv2	Vid-TLDR	
P	C	26.27 / 23.85	29.84 / 35.90	17.76 / 21.24	**35.87** / 40.75	18.07 / 15.28	15.82 / 17.52	23.53 / 30.31	28.87 / **44.47**	26.71 / 22.11	21.33 / 22.61	24.41 / 27.40
	A	25.77 / 41.79	30.47 / 36.48	24.51 / 22.06	**39.19** / 27.03	19.26 / 17.41	25.39 / 20.07	22.70 / 34.16	32.10 / **42.45**	27.07 / 28.86	22.53 / 26.87	26.90 / 29.72
	L	24.77 / 34.62	32.79 / 43.24	17.56 / 19.32	27.79 / 34.03	21.53 / 5.65	15.41 / 16.05	29.19 / 34.77	**37.54 / 45.32**	28.52 / 25.56	24.07 / 27.04	25.92 / 28.56
	CA	30.77 / 30.84	29.42 / 31.95	16.07 / 22.03	28.38 / 33.29	22.71 / 14.32	15.07 / 15.79	23.78 / 27.29	**33.27 / 35.37**	25.53 / 28.40	24.38 / 23.75	24.94 / 26.30
	CL	27.43 / 27.41	31.05 / **42.27**	19.00 / 18.99	**34.96** / 31.24	14.46 / 5.51	16.73 / 12.10	27.34 / 36.75	32.55 / 37.67	28.16 / 27.72	21.26 / 24.50	25.29 / 26.42
	AL	31.58 / 40.70	29.77 / **47.51**	21.95 / 22.76	34.39 / 31.56	18.46 / 6.48	24.25 / 18.94	24.71 / 38.70	**34.80** / 45.68	28.96 / 29.07	25.34 / 27.24	27.42 / 30.86
	CAL	30.74 / 27.62	29.94 / **41.77**	19.86 / 22.31	31.88 / 37.14	19.68 / 7.35	16.12 / 13.20	25.36 / 29.93	**35.01** / 37.14	28.36 / 30.20	25.67 / 23.27	26.26 / 26.99
I	C	38.62 / 30.03	49.16 / 42.90	23.24 / 23.43	36.10 / 47.36	31.79 / 13.53	25.40 / 22.77	39.02 / 43.56	**52.24 / 57.43**	28.83 / 31.52	25.36 / 26.73	34.98 / 33.93
	A	57.69 / 72.14	54.67 / 50.88	25.60 / 27.57	34.50 / 25.81	33.46 / 27.86	33.72 / 42.08	55.86 / 76.10	**65.57 / 82.26**	26.94 / 24.05	29.81 / 35.63	41.78 / 46.44
	L	56.91 / 69.04	59.62 / 55.88	26.68 / 32.04	32.94 / 32.20	41.81 / 24.15	32.91 / 44.12	64.32 / 74.46	**73.03 / 86.22**	29.81 / 30.50	33.27 / 42.57	45.13 / 49.12
	CA	42.23 / 46.37	43.93 / 47.72	20.89 / 26.45	32.48 / 34.32	33.23 / 15.87	26.04 / 27.55	38.00 / 44.53	**52.11 / 61.38**	30.44 / 33.21	27.06 / 30.75	34.64 / 36.81
	CL	39.28 / 56.08	47.29/ 55.57	26.33 / 34.80	32.40 / 34.63	15.64 / 18.75	27.19 / 43.24	44.22 / 67.91	**49.38 / 75.84**	39.23 / 37.16	26.76 / 40.88	34.77 / 46.49
	AL	50.32 / 65.86	49.98 / 45.06	25.63 / 30.33	35.21 / 29.29	26.90 / 22.53	29.27 / 36.92	48.92 / 66.72	**59.48 / 74.18**	32.84 / 25.13	27.32 / 44.02	38.59 / 44.00
	CAL	51.24 / 54.95	51.24 / 39.34	24.70 / 28.83	33.17 / 29.58	29.90 / 18.32	34.37 / 33.18	53.44 / 53.90	**64.27 / 63.81**	33.69 / 26.73	31.85 / 35.74	40.79 / 38.44
AVG		38.48 / 43.62	41.02 / 43.69	22.03 / 24.96	33.29 / 33.58	25.54 / 15.12	24.18 / 25.48	37.68 / 46.17	**47.12 / 55.64**	29.46 / 28.50	26.32 / 30.40	32.51 / 34.72

Table 4: The accuracy (%) of 10 SOTA models on FriendsQA. The figures before the "/" denote single-episode questions, while those after it denote cross-episode questions. Here the bold indicates the best value. Att. means attribution, Top. means topic.

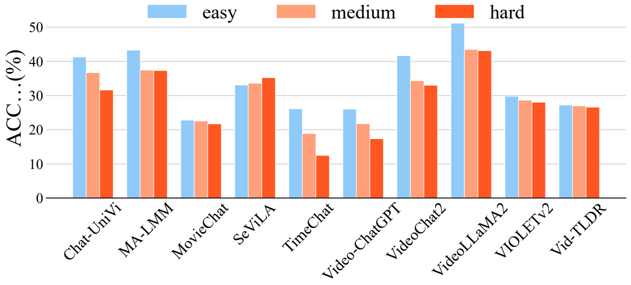


Figure 11: Accuracy (%) of 10 SOTA models across different levels of difficulty in FriendsQA.

ity. Furthermore, extensive language pre-training may have brought more prior knowledge of reasoning to MLLM.

Attribution As shown in Table 4 that the accuracy of P questions is significantly lower than I questions for each model. We calculate the length difficulty σ_l and content difficulty σ_c of both question types. Figure 10(a) shows that P questions get higher overall difficulty, consistent with the results. Specific to 2 factors, although σ_c is similar, P questions have significantly higher σ_l. This suggests the shorter video related to P questions make it harder for the model to extract relevant information.

Topic By Comparing the accuracy of C, A, L within P and I in Table 4, We find that models exhibit lower accuracy on questions involving C. Furthermore, this trend is reinforced by comparing the accuracy of CA, CL, and AL topics within the P. This issue may stems from the model's weakness in character recognition. Unlike traditional facial recognition, in this experiment, we do not provide an additional facial registration database. The model must infer characters by integrating visual content with subtitles, making character recognition a challenging task in this context. However, this pattern is not entirely consistent for cross-episode I questions. We attribute this exception to the fact that I questions can also be answered by reasoning about actions or locations, without solely relying on character recognition.

Difficulty For questions difficulty, we first examine the impact of the two difficulty factors on model performance. Figure 10(b) shows the average accuracy of 10 models across 3 difficulty levels as the two difficulty scores increase. The downward trend in accuracy as difficulty rises confirms the effectiveness of these factors. Figure 11 shows the accuracy of each models in 3 difficulty levels. As observed, 9 of 10 models are consistent with our expectation, where accuracy decreases as difficulty increases. Only SeViLA shows no significant difference among different difficulty levels. It might be because SeViLA is the only model fine-tuning on TVQA including *Friends*, with more prior knowledge.

Conclusion and Future Work

In this paper, we devise StoryMind, a LLM-based multi-agent collaboration framework to automatically generate large-scale dataset, FriendsQA, with fine-grained topics. FriendsQA builds upon the core C, A, and L story topics of stories with a large number of questions. We conduct comprehensive evaluation on FriendsQA and believe that it will guide the development of VideoQA methods.

Although FriendsQA represents a significant advancement, it remains in early stages of development. We are continually refining and expanding StoryMind framework to enhance its adaptability and applicability to various types of story videos. Inspired by (Wang et al. 2020; Nguyen et al. 2024), future work includes extending StoryMind to other genres and shows with different storytelling styles and wider range of scenarios, such as movies and dramas.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62372339, 62371350, 62372336) and the Ministry of Education Industry-University Cooperative Education Project (No. 240700006245501, 220800006041043, 202102246004, 202002142012). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Ammanabrolu, P.; Cheung, W.; Broniec, W.; and Riedl, M. O. 2021. Automated Storytelling via Causal, Commonsense Plot Ordering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7): 5859–5867.

Chen, M.; Chu, Z.; Wiseman, S.; and Gimpel, K. 2021. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Choi, J.; Lee, S.; Chu, J.; Choi, M.; and Kim, H. J. 2024. vid-TLDR: Training Free Token Merging for Light-weight Video Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18771–18781.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models.

Curtis, K.; Awad, G.; Godil, A.; and Soboroff, I. 2023. The ACM Multimedia 2023 Deep Video Understanding Grand Challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 9606–9609. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Curtis, K.; Awad, G.; Rajput, S.; and Soboroff, I. 2020. HLVU: A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, 355–361. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370875.

Curtis, K.; Awad, G.; Rajput, S.; and Soboroff, I. 2022. The ACM Multimedia 2022 Deep Video Understanding Grand Challenge. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 7075–7078. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Fabian Caba Heilbron, B. G., Victor Escorcia; and Niebles, J. C. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–970.

Fu, T.-J.; Li, L.; Gan, Z.; Lin, K.; Wang, W. Y.; Wang, L.; and Liu, Z. 2023. An Empirical Study of End-to-End Video-Language Transformers With Masked Visual Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22898–22909.

Fung, Y.; Wang, H.; Wang, T.; Kebarighotbi, A.; Bansal, M.; Ji, H.; and Natarajan, P. 2023. DeepMaven: Deep question answering on long-distance movie/TV show videos with multimedia knowledge extraction and synthesis. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3041–3051.

Garcia, N.; Otani, M.; Chu, C.; and Nakashima, Y. 2020. KnowIT VQA: Answering Knowledge-Based Questions about Videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07): 10826–10834.

Guo, J.; Liang, C.; and Wang, Z. 2023. Who, What and Where: Composite-semantic Instance Search for Story Videos. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 858–863. IEEE.

He, B.; Li, H.; Jang, Y. K.; Jia, M.; Cao, X.; Shah, A.; Shrivastava, A.; and Lim, S.-N. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13504–13514.

Jin, P.; Takanobu, R.; Zhang, W.; Cao, X.; and Yuan, L. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13700–13710.

Kim, K.-M.; Heo, M.-O.; Choi, S.-H.; and Zhang, B.-T. 2017. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. TVQA: Localized, Compositional Video Question Answering. In *Empirical Methods in Natural Language Processing*.

Lei, J.; Yu, L.; Berg, T.; and Bansal, M. 2020. TVQA+: Spatio-Temporal Grounding for Video Question Answering. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8211–8225. Online: Association for Computational Linguistics.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; Wang, L.; and Qiao, Y. 2024. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22195–22206.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2046–2065. Online: Association for Computational Linguistics.

Li, R.; Guo, J.; Li, M.; Wu, Z.; and Liang, C. 2023. A Hierarchical Deep Video Understanding Method with Shot-Based Instance Search and Large Language Model. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 9425–9429. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Liang, C.; Zhang, Y.; Cheng, J.; Xu, C.; and Lu, H. 2009. A Novel Role-Based Movie Scene Segmentation Method. In Muneesawang, P.; Wu, F.; Kumazawa, I.; Roeksabutr, A.; Liao, M.; and Tang, X., eds., *Advances in Multimedia Information Processing - PCM 2009*, 917–922. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-10467-1.

Ma, K.; Jurczyk, T.; and Choi, J. D. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2039–2048.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 46212–46244. Curran Associates, Inc.

Nguyen, T.; Hu, Z.; Wu, X.; Nguyen, C.-D.; Ng, S.-K.; and Luu, A. T. 2024. Encoding and Controlling Global Semantics for Long-form Video Question Answering. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7049–7066. Miami, Florida, USA: Association for Computational Linguistics.

Niu, Y.; Liang, C.; Lu, A.; Huang, B.; Wang, Z.; and Guo, J. 2023. Person-action Instance Search in Story Videos: An Experimental Study. *ACM Trans. Inf. Syst.*, 42(2).

Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14313–14323.

Sang, J.; Liang, C.; Xu, C.; and Cheng, J. 2011. Robust movie character identification and the sensitivity analysis. In *2011 IEEE International Conference on Multimedia and Expo*, 1–6.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4631–4640.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, X.; Chen, J.; Wang, Z.; Liu, W.; Satoh, S.; Liang, C.; and Lin, C.-W. 2020. When Pedestrian Detection Meets Nighttime Surveillance: A New Benchmark. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 509–515. International Joint Conferences on Artificial Intelligence Organization. Main track.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 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 (CVPR)*, 9777–9786.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *ACM Multimedia*.

Xu, Y.; Wei, Y.; and Wu, B. 2023. Query-aware Long Video Localization and Relation Discrimination for Deep Video Understanding. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 9591–9595. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. In *ICCV*.

Yang, Z.; and Choi, J. D. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 188–197.

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2023. Self-Chained Image-Language Model for Video Localization and Question Answering. In *NeurIPS*.

Yu, Y.; Kim, J.; and Kim, G. 2018. A Joint Sequence Fusion Model for Video Question Answering and Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 9127–9134.

Zhang, B.; Fang, Y.; Ren, T.; and Wu, G. 2022. Multimodal analysis for deep video understanding with video language transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7165–7169.