Appendix

This appendix provides additional information not described in the paper due to the page limit. Specifically, section A contains the specific method and more quantitative results of fine-grained topic categorization. Section B discusses more details about the implementation of StoryMind, including the generator and two reviewers. Section C illustrates samples of FriendsQA. Section D represents the quantity distribution of fine-grained topic questions in FriendsQA. For section E, we elaborate on the experiment settings of the evaluation of 10 SOTA models.

A. Fine-Grained Topic Categorization

We employ Gemini 1.5 Pro to categorize fine-grained topics of questions of 5 classic DVU datasets. The prompt template for assessment is shown in Figure 3.

As shown in Table 1, results of fine-grained topic categorization indicate that the current DVU datasets are not evenly distributed. In contrast, FriendsQA has a more comprehensive and balanced distribution of various topics.

B. Details for StoryMind

In this section, we discuss more details about the implementation of StoryMind, including the prompt for the generator and reviewers.

B.1 Prompt for Generator

We use Gemini 1.5 Pro[1], paired with LangChain[2], as the LLM to generate questions for the given story videos. We develop a formatting tool using Langchain's tools module, which enables the generator to output questions and related information according to our specifications. The prompt template for question generation is shown in Figure 4.

The prompt template contains three main parts: video information, description of the fine-grained topic and question example. The first two parts have been discussed in detail in the paper. As for the final question example, we design heuristic question examples for each fine-grained topic, which enables the generator to mimic and generate new questions. Some question examples are presented in Table 2.

B.2 Prompt for Reveiwers

The structure of the prompt template for the reviewer is shown in Figure 5. In this prompt, each row of the CSV file corresponds to a generated question and its choices.

The first step in the reviewer's verification process is the relevance review. This step requires utilizing the video information, question, and choices provided in the prompt. The video information offers a textual description of the video content, allowing reviewers to determine if the generated questions and choices are relevant to the video (Figure 6(a) in the paper).

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

[2]https://www.langchain.com/

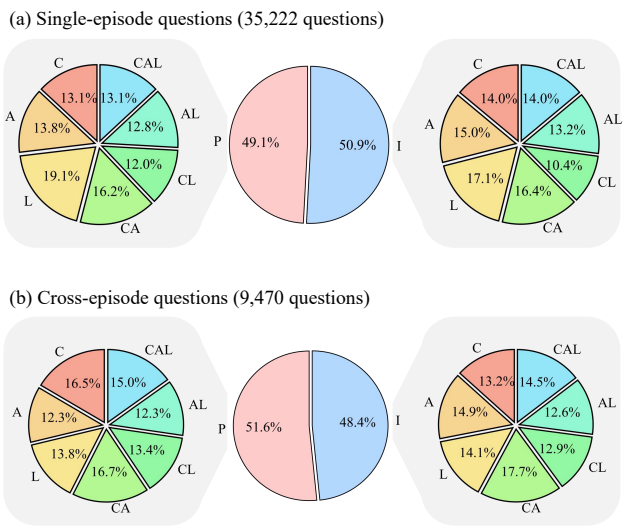

Figure 1: Quantity distributions of single-episode and cross-episode questions of FriendsQA in fine-grained topics

The second step in the reviewer's verification process is the correctness review. A generated question passes the review if and only if the answers provided by two reviewers are consistent with the GT (Figure 6(b) in the paper).

C. Samples of FriendsQA

Figure 6 and Figure 7 show several single-episode and cross-episode questions in FriendsQA, respectively. For easy verification, we specifically identify each question's fine-grained topic category and difficulty score.

Particularly, for the proposed difficulty measure, we give some samples for different difficulty levels in Figure 8 to illustrate how time and content factors affect the difficulty measure of the question. For the time factor, the difficulty score and level increase as related time decreases. Similarly, for the content factor, both the difficulty score and level increase as the number of related instances decreases.

D. Quantity Distributions of FriendsQA

As illustrated in Figure 1, the quantity distribution of FriendsQA covers all fine-grained topics and is balanced across them, including single-episode and cross-episode questions.

E. Setting for Evaluation

The setting for evaluation includes experimental details and the text input of evaluation on FriendsQA.

E.1 Experimental Details

Following recent VideoQA benchmark (Mangalam, Akshulakov, and Malik 2023; Li et al. 2024), we adopt a zero-shot question-answering setting on FriendsQA to test 10 state-of-the-art (SOTA) VideoQA models. To ensure the objectivity of the evaluation, we do not modify the model code and solely utilize the default parameters provided by the official repository for performance evaluation.

Table 1: Categorization results of 14 fine-grained topic categorizations in different DVU datasets.

Topic	MovieQA		TVQA		TVQA+		DVU22&23		MovieChat-1K		**FriendsQA**	
	P	I	P	I	P	I	P	I	P	I	P	I
C	4,097	2,528	31,412	11,313	8,464	3,292	2	48	4,840	384	3,074	3,110
A	85	48	16,513	1,049	1,349	62	44	3	627	2	2,986	3,369
L	208	92	5,311	199	848	16	0	0	8,836	413	3,976	3,712
CA	3,157	1,746	45,567	8,934	9,356	1,099	34	179	1,031	5	3,618	3,747
CL	204	42	8,432	349	1,635	19	2	38	42	0	2,727	2,453
AL	7	2	462	42	70	5	0	3	7	0	2,812	2,934
CAL	1,792	936	7,813	7,517	2,011	1,157	13	89	2,060	770	3,006	3,168

(a) Question Input

Please watch the video carefully and focus on understanding three elements of the video which are character, action, and location. Note that the information about the scene that appears in the question is incrementally labeled according to the location switching. Please answer the following question according to in-depth and comprehensive understanding of the video:

Question: `[question]`

(b) Prompt for MLLM

`[Question Input]`
Please select the best option from following choices:
`[choices]`

Figure 2: Question input and the prompt for MLLM.

E.2 Text Input of Evaluation

This paper evaluates two types of VideoQA models, VLM and MLLM. The similarity between them during evaluation is that they both use video keyframes as the input of visual information, but the difference lies in the input of text information. For VLM, we input question and choices into the question input head and choice input head of the model, respectively.

For MLLM, we put the question and choices as a whole into the prompt and input them into the model (Figure 2). It should be emphasized that in addition to the main question, we also write a description in the question input part to prompt the model to accurately locate the question scene based on the change of locations.

F. Qualitative Experiments

Figure 9 and Figure 10 present qualitative results on the FriendsQA dataset for 10 SOTA models, encompassing both single-episode and cross-episode questions.

References

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.

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.

Table 2: Question examples of the prompt for question generation.

Episode	Attribution	Topic	Examples
Single	P	C	How many important characters from within the TV series appear in Scene 1?
			In Scene 2, in what order do the characters appear?
		A	What is the main action taking place in Scene 2?
			What is the order of appearance of actions in Scene 4?
		L	In which location does Scene 5 take place?
			In Scene 3, where did the event take place?
		CA	What action does Chandler perform while Joey is practicing his lines?
			Which character takes a puff of a cigarette during the rehearsal scene?
		CL	In Scene 13 which characters are in Central Perk?
			In which location do Monica and Ross appear together?
		AL	What action is taking place in Scene 3 at the hospital?
			What sequence of actions takes place in Scene 7 at Central Perk?
		CAL	Who is discussing a topic at Monica's apartment in Scene 12?
			In Scene 3 at Central Perk, who is sitting, discussing, and then standing up?
	I	C	Who are the main characters in this video that contribute to the plot?
			Across the entire episode, how does Rachel's (growth/moods change/opinions)?
		A	In both Scene 1 and Scene 3, what actions appear simultaneously in both scenes?
			How do the actions related to preparing and consuming meals highlight the friendship throughout the episode?
		L	Throughout the episode, how does Monica's apartment serve as a place for?
			How do the scenes set at Central Perk contribute to the development of the friendships over the course of the episode?
		CA	In Scene 2, after Joey greets everyone, what action do the other characters perform?
			How do Ross's interactions and actions with Monica and Phoebe in various scenes depict her role within the friend group?
		CL	Which characters appear at Chandler's apartment in Scene 4 and also appear at Central Perk in Scene 14?
			Which characters appear in both Monica and Rachel's apartment in Scene 12 and Central Perk in Scene 5?
		AL	What actions happen in locations of Scene 12 and Scene 15?
			Across multiple scenes at Central Perk, what actions reflect the reliance on this location for emotional support?
		CAL	In Scene 6 at the street, who gives $1000 and a football phone, and what action do they perform next?
			How do the interactions at Central Perk and Monica's apartment across various scenes reveal the core dynamics of the friends group?
Cross	P	C	Which two characters are shown to be brother and sister?
			In these episodes, which two characters fall in love with each other?
		A	What recurring action is performed whenever someone enters their apartment?
			What action is considered a symbolic gesture of leaving her old life of depending on her family?
		L	How many locations appears in these episodes?
			Which location appears most in these episodes?
		CA	What action does Ross repeatedly perform as he reminisces about his relationship with Carol?
			Which character is the most likely to use humor as a defense mechanism?
		CL	Which character is shown his/her work in Central Perk throughout all episodes?
			In which locations does Ross reveal his past with Carol?
		AL	What location is frequently used for watching TV?
			What action is repeatedly performed in both Monica's apartment and Central Perk?
		CAL	Who is shown working at Central Perk after giving up their credit cards?
			At which locations does Chandler smoke a cigarette?
	I	C	Which two characters are shown to be twins or brother and sister?
			Who appears to be the most pragmatic and realistic in their approach to life?
		A	How do the reactions to receiving unexpected gifts reveal the beliefs?
			How does the recurring action of 'hugging' demonstrate the importance of physical touch and intimacy in their friendships?
		L	How do the different apartments featured in the episodes reflect the personalities and lifestyles?
			How does the setting of Central Perk contribute to the overall tone and atmosphere of the episodes?
		CA	Based on their interactions, which character seems to be the most empathetic and understanding of others' feelings?
			Why does Monica keep her relationship with Alan a secret from her friends for so long?
		CL	How does the contrast between Monica and Rachel's apartment and Ross's apartment reflect their differing emotional states?
			How does the setting of Monica and Rachel's apartment reflect the evolving dynamics of their friendship?
		AL	How does the location of the office contribute to the portrayal of professional life and personal struggles?
			At which location frequently discuss romantic relationships and drinking coffee?
		CAL	Which character works in a restaurant but dreams of having their own restaurant?
			Why is Monica's apartment the primary location for the group's emotional discussions and support?

Prompt for the Categorization of Fine-grained Topic

System: You are an expert in long story video comprehension and you are good at annotate questions' type of DVU task. You need to annotate given question with choices list according to the descriptions of attribution and topic, categorize questions into correct attribution and topic.

Question for story video comprehension is divided into 2 attributions, 7 topics.
Attributions: P, I.
Description: For questions of perception (P), it can be obtained from the appearance of the video directly, while questions of inference (I) need to analyze the content of the video and logical reasoning to get the results.
Topics: C, A, L, CA, CL, AL, CAL
Description: We focus on three core elements of the video: character, action, and location, where C stands for character, A stands for action, and L stands for location. Therefore, there are 7 possible topics, i.e., C, A, L, CA, CL, AL, CAL. It should be noted that topic is based on which elements it contains. For example, C can only involve character, not anything about action or location, questions with topic C mean it's only about character itself.

For example:
"Why does Hook leave Eamon's apartment?" contains character ('Hook'), action ('leave') and location ('Eamon's apartment') and to answer 'Why' needs logical inference. So this question attribution is 'I', topic is 'CAL'.
"Who is helping the Loyalists make the bomb?" contains character ('Who'), action ('make the bomb') and to answer 'Who' needs 'perception' from video. So this question attribution is 'I', topic is 'CA'.

Questions to be annotated are as follow:
[questions_str]

Output Reuirements
Each line includes 3 elements, separated by spaces: id of the question, attribution of the question (P or I), topic of the question (select from C, A, L, CA, CL, AL, CAL).

Output examples:
0 P CAL
1 I CAL
2 P CAL
3 I CAL
4 P CAL
…

Now you are annotating questions whose id from [start] to [end], please give the annotating result according to the id from [start] to [end] (Directly output without any explanation.):

Figure 3: The prompt template for assessing fine-grained topics of different DVU datasets.

Prompt for Generator

System: You are an expert in long story video comprehension and now need to put your students to the test by coming up with a series of questions for them to answer.

Question for story video comprehension is divided into 2 attributions, 7 topics.
Attributions: P, I.
Description: For questions of perception (P), it can be obtained from the appearance of the video directly, while questions of inference (I) need to analyze the content of the video and logical reasoning to get the results.
Topics: C, A, L, CA, CL, AL, CAL
Description: We focus on three core elements of the video: character, action, and location, where C stands for character, A stands for action, and L stands for location. Therefore, there are 7 possible topics, i.e., C, A, L, CA, CL, AL, CAL. It should be noted that topic is based on which elements it contains. For example, C can only involve character, not anything about action or location, questions with topic C mean it's only about character itself.

Please note that the video information contains the boundingbox, character, and line and line timestamps:
[Video information]

Questions Examples:
[Questions Examples]

Requirements: I hope relevant segments can be cross video clips. I want to generate questions of different difficulty, so for each type you need to help me generate questions of different difficulty by either focusing on more long-term video clips, or focusing on more complex character's relationship. You can even generate questions cover the whole episode. And the generation of questions should ideally be spread throughout the video rather than concentrated in certain scenes. In order to better validate the question, Please generate the basis of inference, a list of the start and end times of the relevant segments in the video, and a list of the relevant characters related to question.

Please generate questions with GroundTruth and choices list in multiple-choice format. For each 2 attributions and 7 topics, i.e., for each of the 14 type, you should generate at least 15 questions and questions should be with different degree of difficulty. You are specialize in using ‘Tools_of_saving_question’ to save questions and you can generate about 14 questions and save. Use tool ‘Tools_of_saving_question’ to save question information.

Figure 4: The prompt template for generator.

Prompt for Reviewer

System: You are very good at reviewing questions for correctness and answering given questions. Here is a video information for a movie or TV show, and the corresponding test questions for understanding the content of the movie or TV show. You are asked to evaluate each question, assessing its correctness (True or False). Assuming you are actually doing the test and can only watch the video, you only need to give the corresponding assessment in the order in which they are presented.

Please note that the video information contains the boundingbox, character, and line and line timestamps:
[Video information]

Generated questions are as follow:
[CSV]

Relevant Requirements:
For correctness, There are two criteria for the correctness of a question and both of them should be satisified:
1. The question must be relevant to the content of the video and can be answered with the only one Answer from Choices.
2. In the list of options for this question, there must be only one correct answer, with the other options unequivocally incorrect.
You need to carefully review each question, ensuring that there are no factual errors, logical reasoning errors, etc., in the question and answer. For conservative, Please filter out the 20% of questions with low confidence in all questions by giving their correctness as False.

Output Requirements:
Each line includes 3 elements, separated by spaces: id of the question, correctness of the question (True or False), the answer you want to use to answer this question (You can select from Choices.).\\

Output examples:
1 True XXX
2 True XXX
3 True XXX
4 True XXX
5 False XXX
...

Figure 5: The prompt template for reviewer.

Question: How many people are present in the first scene of the episode?

✗ (A) 5
✓ (B) 6
✗ (C) 7
✗ (D) 8

Attribution	Topic	Difficulty
P	C	1.6558

Question: What is the main action taking place in the first scene of the episode?

✓ (A) The friends are discussing the importance of kissing
✗ (B) Ross is setting up a museum exhibit
✗ (C) Rachel is looking for her engagement ring
✗ (D) Monica is cooking dinner for her parents

Attribution	Topic	Difficulty
P	A	0.5112

Question: In which location does Scene 9 take place?

✗ (A) Central Per
✗ (B) Museum of Prehistoric History
✓ (C) Barry's office
✗ (D) Monica and Rachel's

Attribution	Topic	Difficulty
P	L	0.9343

Question: Who is worried about the 'karmic debt' of keeping the extra money?

✓ (A) Phoebe
✗ (B) Chandler
✗ (C) Rachel
✗ (D) Monica

Attribution	Topic	Difficulty
I	C	2.8298

Question: In both Scene 1 and Scene 3, what is the main topic of conversation?

✗ (A) The ethics of keeping money that isn't yours
✓ (B) The challenges of dating and relationships
✗ (C) The importance of honesty and integrity
✗ (D) The nature of true friendship and support

Attribution	Topic	Difficulty
I	A	2.9279

Question: Which scenes take place at Central Perk?

✗ (A) Scene 1, Scene 3, and Scene 5
✓ (B) Scene 1, Scene 5, and Scene 9
✗ (C) Scene 2, Scene 4, and Scene 6
✗ (D) Scene 4, Scene 8, and Scene 12

Attribution	Topic	Difficulty
I	L	0.2211

Question: In Scene 13, what is Angela doing while Monica is telling the story about Underdog?

✗ (A) Talking to Monica
✗ (B) Looking at Joey
✗ (C) Drinking wine
✓ (D) Touching Bob's arm

Attribution	Topic	Difficulty
P	C A	1.1676

Question: In Scene 10, where are Monica and Angela?

✗ (A) At Central Perk
✗ (B) At the Launderama
✗ (C) In the kitchen
✓ (D) In the ladies' bathroom at the restaurant

Attribution	Topic	Difficulty
P	C L	1.4311

Question: What action is taking place at the laundromat in Scene 7?

✗ (A) Rachel is doing laundry with Ross's help.
✓ (B) Rachel is arguing with a woman over a laundry machine.
✗ (C) Rachel is sorting her pink clothes.
✗ (D) Rachel is waiting for Ross to arrive at the laundromat.

Attribution	Topic	Difficulty
P	A L	2.2159

Question: In Scene 15 at the laundromat, what is Ross's reaction when the woman tries to take the laundry cart after Rachel gets upset?

✗ (A) He laughs at Rachel's reaction.
✓ (B) He defends Rachel and argues with the woman.
✗ (C) He ignores the situation.
✗ (D) He tells Rachel to let the woman take the cart.

Attribution	Topic	Difficulty
P	C A L	1.8753

Question: Why is Chandler not joining the Thanksgiving festivities in Scene 8?

✗ (A) He doesn't like turkey.
✓ (B) He has bad memories of Thanksgiving.
✗ (C) He's going out of town.
✗ (D) He's fighting with Monica.

Attribution	Topic	Difficulty
I	C A	0.9889

Question: In Scene 10 and Scene 12, which characters are present at both Monica and Rachel's apartment and the hallway during the key incident?

✓ (A) Monica, Rachel, Phoebe, Joey, and Chandler
✗ (B) Monica, Rachel, Ross, Joey, and Chandler
✗ (C) Monica, Phoebe, Ross, Joey, and Chandler
✗ (D) Rachel, Phoebe, Ross, Joey, and Chandler

Attribution	Topic	Difficulty
I	C L	0.3915

Question: How do the actions taking place in both Monica and Rachel's apartment and Carol and Susan's apartment throughout the episode illustrate the theme of family and chosen family?

✗ (A) Sharing meals, expressing gratitude
✗ (B) Offering support, providing a sense of belonging
✗ (C) Celebrating holidays, upholding traditions
✓ (D) All of the above

Attribution	Topic	Difficulty
I	A L	0.2161

Question: In Scene 2 at Central Perk, who is talking about their parents' Thanksgiving plans, and what action do they perform next?

✓ (A) Rachel, She takes a sip of her coffee
✗ (B) Ross, He gets up to call their mother
✗ (C) Monica, She starts preparing dinner
✗ (D) Joey, He makes a joke about the situation

Attribution	Topic	Difficulty
I	C A L	1.0768

Figure 6: Examples of the fine-grained topic questions (single-episode questions)

S01E01

S01E02

S01E03

S01E04

Question: What is the significance of October 20th for Ross?
✗(A) It's the anniversary of his first date with Carol.
✗(B) It's the anniversary of his parents' wedding.
✓(C) It's the anniversary of his first time having sex.
✗(D) It's the anniversary of his divorce from Carol.

Attribution	Topic	Difficulty
P	C	3.8149

Question: Why does Monica feel uncomfortable with her friends' overwhelming enthusiasm for Alan?
✗(A) She feels like they are being too critical of him.
✗(B) She is worried that they will scare him away.
✓(C) She is concerned that their positive opinions will influence her own judgment.
✗(D) She is jealous of the attention they are giving him.

Attribution	Topic	Difficulty
I	C	1.1676

Question: What action does the character perform as a symbolic gesture of leaving her old life of depending on her father?
✗(A) She moves out of her parents' house.
✗(B) She gets a job and starts supporting herself.
✓(C) She cuts up her credit cards.
✗(D) She returns her engagement ring to Barry.

Attribution	Topic	Difficulty
P	A	2.0000

Question: How does the recurring action of watching television together reflect the characters' need for escapism and shared experiences?
✗(A) It allows them to avoid confronting their own problems and anxieties.
✗(B) It provides a source of entertainment and distraction from their daily lives.
✗(C) It fosters a sense of community and shared cultural references.
✓(D) All of the above.

Attribution	Topic	Difficulty
I	A	1.1145

Question: Which location is seen the least throughout the episodes?
✗(A) Monica and Rachel's apartment
✗(B) Chandler and Joey's apartment
✗(C) Ross's apartment
✓(D) The Lamaze class

Attribution	Topic	Difficulty
P	L	0.8421

Question: How does Central Perk, as a setting, reflect the changing dynamics and emotional journeys of the characters throughout the episodes?
✗(A) It serves as a neutral ground where the characters can come together to celebrate, commiserate, and support each other.
✗(B) It reflects the characters' evolving social lives and their interactions with the wider world.
✗(C) It represents a constant in their lives amidst personal and relational changes.
✓(D) All of the above

Attribution	Topic	Difficulty
I	L	1.0927

Question: Who expresses a strong desire to be married again after their divorce?
✓(A) Ross
✗(B) Monica
✗(C) Chandler
✗(D) Rachel

Attribution	Topic	Difficulty
P	C A	14.9104

Question: Based on their interactions, which character demonstrates the most empathy and understanding towards others' emotions?
✗(A) Monica
✗(B) Rachel
✓(C) Phoebe
✗(D) Joey

Attribution	Topic	Difficulty
I	C A	5.9441

Question:At which location does Rachel admit to Monica that she's attracted to Ross?
✗(A) Central Perk
✓(B) Monica and Rachel's apartment
✗(C) Chandler and Joey's apartment
✗(D) The hospital

Attribution	Topic	Difficulty
P	C L	6.6107

Question: How does Rachel's experience at Barry's office impact her perspective on her past relationship and her future aspirations?
✗(A) It reinforces her decision to leave Barry and pursue a new path.
✗(B) It highlights the stark contrast between her old life and her new aspirations.
✗(C) It triggers a sense of nostalgia and regret for what she left behind.
✓(D) All of the above

Attribution	Topic	Difficulty
I	C L	3.3293

Question:What action is repeatedly performed at Central Perk?
✓(A) Ordering coffee
✗(B) Playing foosball
✗(C) Watching television
✗(D) Reading books

Attribution	Topic	Difficulty
P	A L	0.8458

Question: Why do the characters frequently gather at Monica and Rachel's apartment to watch television and discuss their lives?
✗(A) It provides a comfortable and familiar space for them to relax and unwind.
✗(B) It allows them to bond over shared experiences and pop culture references.
✗(C) It fosters a sense of intimacy and connection as they discuss their personal lives.
✓(D) All of the above

Attribution	Topic	Difficulty
I	A L	1.1145

Question: Who gives a homeless person $1000 and a football phone on the street?
✓(A) Phoebe
✗(B) Chandler
✗(C) Ross.
✗(D) Monica.

Attribution	Topic	Difficulty
P	C A L	4.0688

Question: Which character works in a restaurant but dreams of having their own restaurant?
✓(A) Monica works as a chef in a restaurant but dreams of opening her own restaurant one day.
✗(B) Joey works as a waiter at Central Perk but dreams of becoming a successful actor.
✗(C) Rachel works as a waitress at Central Perk but dreams of a career in fashion.
✗(D) Phoebe works as a masseuse but dreams of becoming a professional musician.

Attribution	Topic	Difficulty
I	C A L	7.1029

Figure 7: Examples of the fine-grained topic questions (cross-episode questions)

Figure 8: Samples of FriendsQA in different difficulty levels

S01E01

Question: How does Rachel's demeanor change from the beginning of the episode to the end? ***Difficulty:*** *0.8390* ***Level:*** *Easy*

✗(A) She becomes more dependent on her parents.
✗(B) She remains optimistic and cheerful throughout.
✓(C) She transitions from being reliant to seeking independence.
✗(D) She becomes increasingly pessimistic about her future.

Chat-UniVi:
(C) She transitions from being reliant to seeking independence.

Video-ChatGPT:
Rachel's demeanor changes from being optimistic and cheerful throughout to becoming increasingly pessimistic about her future.

MA-LMM:
C) She transitions from being reliant to seeking independence.

VideoChat2:
C) She transitions from being reliant to seeking independence.

MovieChat:
(D) She becomes increasingly pessimistic about her future.

VideoLLaMA2:
Answer: (D) She becomes increasingly pessimistic about her future.

SeViLA:
She becomes more dependent on her parents.

VIOLETv2:
She transitions from being reliant to seeking independence.

TimeChat:
The given query happens in 1195.0 - 1227.5 seconds.

Vid-TLDR:
She transitions from being reliant to seeking independence.

Question: What is Phoebe doing at the beginning of Scene 4? ***Difficulty:*** *4.7126* ***Level:*** *Medium*

✓ (A) Playing guitar and singing for money
✗(B) Working at Central Perk
✗ (C) Shopping for clothes
✗(D) Visiting her grandmother

Chat-UniVi:
(A) Playing guitar and singing for money.

Video-ChatGPT:
Phoebe is working at Central Perk.

MA-LMM:
A) Playing guitar and singing for money.

VideoChat2:
C) Shopping for clothes.

MovieChat:
A) Playing guitar and singing for money.

VideoLLaMA2:
Answer: (B) Working at Central Perk

SeViLA:
Working at Central Perk

VIOLETv2:
Shopping for clothes

TimeChat:
(B) Working at Central Perk.

Vid-TLDR:
Working at Central Perk

Question: Where does Monica work? ***Difficulty:*** *13.5915* ***Level:*** *Hard*

✗ (A) Central Perk ✗ (B) A Restaurant ✓(C) Iridium ✗(D) A fashion boutique

Chat-UniVi:
(A) Central Perk.

Video-ChatGPT:
A museum.

MA-LMM:
(A) Central Perk.

VideoChat2:
D) A restaurant

MovieChat:
(A) Central Perk.

VideoLLaMA2:
Answer: (A) Central Perk

SeViLA:
Central Perk

VIOLETv2:
Central Perk

TimeChat:
(D) A restaurant.

Vid-TLDR:
Iridium

Figure 9: Qualitative results across 10 SOTA models on single-episode questions.

S01E01-S01E04

Question: How does Rachel's experience with her friends' reactions to her previous boyfriends influence her perception of Alan?
✓(A) She is skeptical of their judgment and worries they will not approve of Alan.
✗(B) She is eager for their approval and wants them to like Alan as much as she does.
✗(C) She is indifferent to their opinions and focuses solely on her own feelings.
✗(D) She is afraid that their negative experiences will jinx her relationship.

Difficulty: *2.3238* ***Level:*** *Easy*

Chat-UniVi:
(C) She is indifferent to their opinions and focuses solely on her own feelings.

Video-ChatGPT:
She is eager for their approval and wants them to like Alan as much as she does.

MA-LMM:
A) She is skeptical of their judgment and worries they will not approve of Alan.

VideoChat2:
A) She is skeptical of their judgment and worries they will not approve of Alan.

MovieChat:
A) In the video, Rachel is shown to be interacting with her female friends...

VideoLLaMA2:
Answer: (A) She is skeptical of their judgment and worries they will not approve of Alan.

SeViLA:
79.3 seconds

VIOLETv2:
She is skeptical of their judgment and worries they will not approve of Alan.

TimeChat:
She is eager for their approval and wants them to like Alan as much as she does.

Vid-TLDR:
She is afraid that their negative experiences will jinx her relationship.

Question: How does the recurring motif of 'losing things' in the episodes symbolize the characters' broader anxieties and uncertainties?
✗(A) It reflects their fear of losing control and stability in their lives.
✗(B) It symbolizes their struggle to find meaning and purpose.
✗(C) It highlights their vulnerability to unexpected events and emotional setbacks.
✓(D) All of the above

Difficulty: *4.6518* ***Level:*** *Medium*

Chat-UniVi:
(D) All of the above

Video-ChatGPT:
All of the above. The recurring motif of 'losing things' ...

MA-LMM:
C) It highlights their vulnerability to unexpected events and emotional setbacks.

VideoChat2:
D) All of the above

MovieChat:
A) It reflects their fear of losing control and stability in their lives.

VideoLLaMA2:
Answer: (D) All of the above

SeViLA:
It reflects their fear of losing control and stability in their lives.

VIOLETv2:
It reflects their fear of losing control and stability in their lives.

TimeChat:
It symbolizes their struggle to find meaning and purpose

Vid-TLDR:
It reflects their fear of losing control and stability in their lives.

Question: Why does Rachel say that she feels like she gave up everything for nothing?
✗(A) She misses her old life with Barry
✓(B) She feels lost and uncertain about her future
✗(C) She regrets leaving Barry at the altar
✗(D) She is tired of working as a waitress

Difficulty: *14.6285* ***Level:*** *Hard*

Chat-UniVi:
(C) She regrets leaving Barry at the altar.

Video-ChatGPT:
Rachel feels lost and uncertain about her future.

MA-LMM:
D) She is tired of working as a waitress

VideoChat2:
D) She is tired of working as a waitress

MovieChat:
C) She regrets leaving Barry at the altar

VideoLLaMA2:
Answer: (C) She regrets leaving Barry at the altar

SeViLA:
She regrets leaving Barry at the altar

VIOLETv2:
She misses her old life with Barry

TimeChat:
(B) She feels lost and uncertain about her future.

Vid-TLDR:
She regrets leaving Barry at the altar

Figure 10: Qualitative results across 10 SOTA models on cross-episode questions.