

A HUMAN VERIFICATION

To verify the reliability of our benchmark, we conducted a human evaluation to validate the correctness of our questions and answers pairs. Due to the challenging nature of our benchmark, which involves understanding videos over an hour long, we randomly sample 2.5% of the data for human verification.

More specifically, we asked nine annotators to read the questions related to each video and then watch the full video before validating the questions then start validating the questions. Each video averages 80 questions and is over an hour long, requiring 3-4 hours for annotators to verify. The study shows a great alignment between the human responses and our benchmark, where the accuracy of the true questions across different skills is **92%**. The detailed accuracy of the true questions per skill is reported in the table 1. The remaining two skills, i.e., local visual questions and summarizing,

Skill Name	Number of Questions	Accuracy of Correctness (%)
Character Actions	193	88.08
Deep Context Understanding	600	92.60
Global Appearance	25	80.00
Linking Multiple Events	621	97.42
Scene Transitions	24	87.50
Spoiler Questions	43	95.34
Temporal Questions	804	88.68
Overall accuracy	2310	92.00

Table 1: Human verification for all the GPT skills involved in the generation pipeline

do not need human verification, as the first one is adopted from the TVQA dataset, and the latter is scrapped from human responses on the web.

B EVALUATION MODELS SETTING

In our evaluation, we evaluated two commercial models and four open-source models. We also added the results of two leading short video models to see the performance of short video models.

GPT-4o. GPT-4o cannot natively process .mp4 video files. To work within its limitations, we sampled the maximum of 250 frames from the video, followed by the subtitle text and the question. **Gemini-Flash 1.5.** The Gemini-Flash 1.5 model, developed by Google Gemini (2024), Gemini recently gained the capability to process .mp4 files, but since our benchmark videos lack audio, we provided Gemini with the video file, followed by the subtitle file and the accompanying question. **Goldfish.** we used the defulat model setting with k=3. **LLama-vid.** The LLama-vid model Li et al. (2023) accepts both video frames and subtitles. For our evaluation of the movies, we utilized our dataset with one frame per second, accompanied by aligned subtitle shots. The model was evaluated using the default settings without any modifications to the inference parameters. **Large World Model (LWM).** LWM is efficiently optimized for execution on Google TPUs and has another version for GPUs. Our evaluation is done using (NVIDIA A100), which allows for processing a maximum of 8 frames per video. While this setup does not represent the optimal configuration for LWM, it was the most feasible setting. LWM can accept only the video frames without the subtitles. **Moviechat.** The Moviechat model Song et al. (2023) processes video frames without subtitles and operates in global and breakpoint modes. Our evaluation focused on the global mode, utilizing the default inference settings without any modifications. **LLaVA-NeXT-Interleave.** LLaVA-NeXT-Interleave can process only 8 frames per video, we also used the default setting without any changes. **MiniGPT4-video.** We evaluated MiniGPT4-video with the llama 2 version that capable of handling 45 frames per video.we also used the default setting without any changes.

C EVALUATION DETAILS

C.1 EVALUATION METRIC DETAILS

For MCQs, large language models (LLMs) do not consistently provide direct responses. The output may vary, sometimes giving the option number, other times the option sentence, or occasionally providing additional clarifications for the selected option. For example, an LLM might produce a response such as: "I think option 1 is close, but my final answer will be option 2." Additionally, some responses may include hallucinations not found in the given options. To address this variability, we implemented a standardized evaluation method using GPT-4o to match the LLM's prediction with one of the provided options. Specifically, we input the set of options and the LLM's prediction into GPT-4o, which then attempts to match the predicted answer with one of the given options. If no matching option is found or if the response includes hallucinations, GPT-4o matches the prediction with an "I don't know" option. Using the prediction option number and the ground truth option number, we then calculate the accuracy. For open-ended questions, GPT-4o assessed the LLMs' predictions based on several criteria: correctness, meaningfulness, alignment with the expected answer, presence of hallucinations, and completeness. Using these criteria, GPT-4o assigned a score from 0 to 5 to indicate the overall quality of each response.

C.2 EVALUATION PROMPTS DETAILS

In this section we will discuss the details for the prompts that have been used for evaluation for both the open ended questions and multiple choices. Figure. 1 show the detailed prompt used for the results matching. Figure 2 show the detailed prompt for the GPT-4o scores.

D EXTRA BENCHMARK EXAMPLES

Here in this sections, we are showing more examples of our benchmark skills such as the temporal order of events in Fig. 3, linking multiple events in Figure.4, deep context understanding in Figure. 5 , local questions in Figure.6 ,spoiler questions in Figure.8, Sequence of character actions Figure.9, and summarization in Figure. 7.

E SUCCESS AND FAILURE CASES

In this section, we present examples of both success and failure cases in question generation using GPT-4o. Figure 11 illustrates cases involving the generation of Temporal Order of Events questions, while Figure 10 showcases examples related to Linking Multiple Events questions. As highlighted in the human evaluation section A, such failure cases are infrequent, with 92% of the generated data verified as accurate.

F QUALITATIVE RESULTS

In this section, we present qualitative results to assess how the evaluated models perform in answering the benchmark questions. We also examine how GPT-4o scores these responses compared to the ground truth, particularly in the case of open-ended questions. Figure 12 shows an example of the deep context understanding skill , Figure 13 shows an example of Global appearance skill, Figure 14 shows an example of the Scene transition skill and Figure 15 shows an example of the spoiler questions skill.

in the spoiler questions and deep context understanding , we can see the GPT-4o scores for each answer.

G INFINIBENCH GENERATION DETAILS

This section elaborates on the specific prompts employed to generate questions for each skill category. The prompts, utilized within the GPT-4o framework, are depicted in Figures 16, 18, 17,

19,20. These figures provide the exact phrasing and structure used for question generation, ensuring reproducibility and clarity in the benchmarking creation process.

MCQ matching prompt:

System prompt:

You are an intelligent chatbot designed to evaluate the correctness of generative outputs for multiple-choice questions (MCQs). Your task is to match the predicted answer with one of the provided options, which include an 'I don't know' option. If there is no match between the predicted answer and the options, choose the option that says, 'I don't know'. Here's how you can accomplish the task:

INSTRUCTIONS:

- Focus on finding a meaningful match between the predicted answer and the correct option.
- Consider synonyms or paraphrases as valid matches.
- Choose an option only if you believe there is sufficient evidence to directly derive the answer from the predicted information or indirectly with minimal reasoning. If there isn't enough evidence to support any option, simply select the option with 'I don't know.'
- Provide only the integer that represents the option number for your evaluation decision.
- Evaluate as a human would, considering context and meaning, not just exact words.
- Provide your answer in the form of a Python dictionary string with the key 'decision', such as {'decision': 3}.

User prompt:

Please evaluate the following question-answer pair:

Options: {options}

Predicted Answer: {pred}

Provide your evaluation as a decision with the matched option number.

Generate the response in the form of a Python dictionary string with the key 'decision'.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.

For example, your response should look like this: {'decision': 1}.

Do not include any other information in your response such as ```python```.

Figure 1: Detailed prompt for MCQ evaluation

Scoring evaluation prompt:

System prompt:

You are an intelligent chatbot designed to evaluate the correctness of generative outputs for question-answer pairs.

Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:

INSTRUCTIONS:

- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.
- Provide a score between 0 and 5, where 5 indicates the highest meaningful match.
- Penalize the score if the predicted answer contains hallucinations or is missing key parts of the correct answer.
- Assign your score based on how far the predicted answer is from the correct answer.
- Evaluate as a human would, not as a machine.
- Provide your score in the form of a Python dictionary string with the key 'score', such as {'score': 3.7}.

User prompt:

Please evaluate the following video-based question-answer pair:

Question: {question}

Correct Answer: {answer}

Predicted Answer: {pred}

Provide your evaluation only as a score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match.

Generate the response in the form of a Python dictionary string with the key 'score'.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.

For example, your response should look like this: {'score': 4}.

Do not include any other information in your response such as ```python```.

Figure 2: Detailed prompt for Scoring system evaluation

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

Temporal questions:

Q: Choose the correct option for the following question: Looking at these events : [Chandler reluctantly agrees to return to his old job after negotiation, Monica gets disappointed by the lost job opportunity, Phoebe gives Steve a painful massage as payback, Monica's audition dinner is ruined by Steve being stoned], how do they unfold in the episode?

- Option 1: [Phoebe gives Steve a painful massage as payback ,Chandler reluctantly agrees to return to his old job after negotiation, Monica gets disappointed by the lost job opportunity , Monica's audition dinner is ruined by Steve being stoned],
Option 2: [Monica's audition dinner is ruined by Steve being stoned ,Chandler reluctantly agrees to return to his old job after negotiation, Phoebe gives Steve a painful massage as payback ,Monica gets disappointed by the lost job opportunity],
Option 3: I don't know,
Option 4: [Monica gets disappointed by the lost job opportunity ,Chandler reluctantly agrees to return to his old job after negotiation, Phoebe gives Steve a painful massage as payback , Monica's audition dinner is ruined by Steve being stoned],
Option 5: [Chandler reluctantly agrees to return to his old job after negotiation ,Monica's audition dinner is ruined by Steve being stoned, Monica gets disappointed by the lost job opportunity ,Phoebe gives Steve a painful massage as payback]



Figure 3: Example for the temporal order of events skill

Linking multiple events :

Q: What is the connection between Monica's failed dinner and Phoebe's reaction during Steve's next massage appointment?

Monica's dinner for Steve fails due to his stoned condition and disruptive behavior. Phoebe, out of frustration with Steve's behavior and the ruined dinner, takes out her anger on him during his next massage appointment by giving him a painful massage.

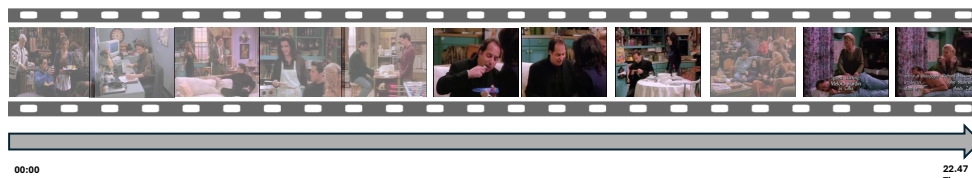


Figure 4: Example for the linking multiple events skill

Deep context understanding:

Q: What does Celia do when Marcel Ross's monkey starts interacting with her during the date?

Celia screams and is unable to handle Marcel pulling at her hair until Ross lifts Marcel away.



Figure 5: Example for the deep context understanding skill

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

Local questions

Q: Choose the correct option for the following question:
What is Joey eating when Chandler is on the phone with the guy from his old job?

- Option 1: A piece of pie
- Option 2: Popcorn
- Option 3: A donut
- Option 4: A bread roll
- Option 5: A slice of pizza



Figure 6: Example for the local questions

Summarization

Q: Please summarize the video with as much detail as possible.

Monica cooks a gourmet meal for Steve (Jon Lovitz), a restaurateur looking for a new head chef. Steve is a massage client for Phoebe, and she makes the introduction between Monica and him. The job is perfect as Steve wants something eclectic and needs someone who can create the entire menu. As an audition, Monica is cooking dinner for him the coming week. She wants Phoebe to be there. Monica hires a professional waitress Wendy (for \$10/hr.), which offends Rachel (Monica says that she needed a professional waitress). Wendy bails on Monica at the last minute. Monica begs Rachel and even says that she gave her shelter when she had nowhere else to go.. Eventually she offers Rachel \$20/hr. He arrives stoned and wants to eat everything in sight, including taco shells and gummy bears. Phoebe tells Rachel who tries to handle the situation by offering Steve some wine. Eventually Monica realizes that Steve is super stoned. She tries to yank the gummy bears from Steve, and they end up falling in the punch bowl.. Dinner is a total disaster, and the gang tells her that she doesn't want to work for a guy like that. After working as a data processor for five years, Chandler gets promoted to supervisor. Chandler quits, claiming he only intended for his job to be temporary (and Chandler already has been there for over 5 yrs.). Chandler goes to meet a career counselor. After 8 hrs. of aptitude, personality and intelligence tests he learns that he is fit for a career in data processing, for a large multinational corporation. he is disappointed as he always pictured himself doing something cool. When his boss calls and offers more money (& more bonus.. Chandler resists, but the boss keeps throwing more and more numbers), Chandler caves and goes back to work. Chandler gets the corner office, and he shows it off to Phoebe. He has a view and an assistant. But Chandler has more responsibility now and starts spending more time & late nights at work and yelling at his juniors. He doesn't like it. Ross has a date with a beautiful colleague named Celia (Melora Hardin) (curator of insects at the museum) and gives new meaning to the term 'spanking the monkey' when she meets Marcel. The date goes bad when Marcel hands on Celia's hair and pulls it. Eventually Ross takes Celia to bed, and she wants him to talk dirty and he says 'Vulva'. Ross turns to Joey for advice as Celia wants him to talk dirty as foreplay. Joey gets Ross to practice on him.. When Ross talks smack, Chandler overhears and amuses himself at their expense. Ross does well at the next date and talks very dirty (with theme, plot, motif and story-lines. at one point there were villagers), but eventually they get tired and cuddle. Phoebe takes out her anger at Steve at his next massage appointment by treating him to a bad massage (she elbows him on his back and pinches his skin so that it hurts).



Figure 7: Example for the summarization skill

Spoiler questions

Q: Why didn't the Arquillian in the jeweler's head simply tell Jay that the galaxy was on his cat's collar?

To add a bit of mystery to the story. If he'd said 'the galaxy in the jewel on the cat's collar', the movie would have ended much faster. Actually, Arquillian was indeed trying to tell Jay that the galaxy was on the cat's collar. He just didn't have the correct vocabulary to do so. Note how he stumbles over the word "war". He almost certainly thinks "belt" is the correct word for "collar", which is understandable because the articles of clothing are identical, as the only differences are that one is worn around the waist and the other is worn around the neck. And the cat's name is Orion, so he's being accurately descriptive, not deceitful. It's likely that the Arquillian didn't understand much English and that the Jeweler's body had a translator in it when conversing with humans. It was likely damaged when Edgar stabbed it through the neck.

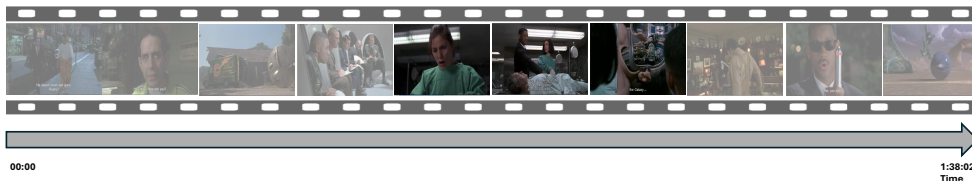


Figure 8: Example for the spoiler questions

Character Actions :

Q: Choose the correct option for the following question: What did Rachel do through this video?

Option 1: enjoying a cappuccino with a dash of cinnamon at a trendy coffee shop , discussing the latest book club read with a friend over lunch , savoring a slice of gourmet pizza with sun-dried tomatoes and arugula at a pizzeria , exploring an art museum's new exhibit on modern sculpture.

Option 2: I don't know

Option 3: Rachel attends an initial interview, accidentally kisses Mr. Zelner. , She calls back, gets another meeting with Mr. Zelner and eventually gets the job after apologizing and explaining her actions. , She practices handshaking with Phoebe and Monica., Accidentally touches Mr. Zelner's crotch while offering a handshake., Rachel gets a second interview call and gets ink on her lips during the second attempt., Rachel enters Central Perk and announces her job interview at Ralph Lauren., Rachel goes home, realizes her mistake regarding the ink., Misinterprets Mr. Zelner's gesture and walks out thinking he's making an advance.

Option 4: Rachel enters Central Perk and announces her job interview at Ralph Lauren., She practices handshaking with Phoebe and Monica. , Rachel attends an initial interview, accidentally kisses Mr. Zelner. , Rachel gets a second interview call and gets ink on her lips during the second attempt. , Misinterprets Mr. Zelner's gesture and walks out thinking he's making an advance. , Rachel goes home, realizes her mistake regarding the ink. , She calls back, gets another meeting with Mr. Zelner and eventually gets the job after apologizing and explaining her actions. , Accidentally touches Mr. Zelner's crotch while offering a handshake.



Figure 9: Example for sequence of character actions questions

Linking multiple events



How does Dr. House's internal conflict towards the end connect to the events of the episode?

Dr. House's internal conflict at the end ties together the various events of the episode. The stress of the almost Sci-fi case, the emotional impact of Foreman's departure, and the unresolved medical mystery all contribute to House's turmoil, leaving him with an immense conflict that sets the stage for the next season.



In what ways do the sea rescue and the medical mystery serve as catalysts for character development within the Diagnostics team?

The sea rescue brings the couple to the team's attention, setting off a series of events that act as catalysts for character development. The challenging medical mystery forces team members to confront their own abilities, resolve conflicts, and cope with Foreman's departure, leading to significant personal and professional growth.




Why the answer is not valid ?

Because the couple traveled a great distance, time and danger to reach the hospital through the Coast Guards. This encourages the House team to do their best for this case.

Figure 10: Examples of success and failure cases in Linking Multiple Events questions.

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431


Temporal questions



00:00 22:05 Time

Choose the correct option for the following question: Given the events listed ['Rachel apologizes to Ross and suggests a romantic dinner to make it up to him', 'Monica goes to her eye appointment with Dr. Burke', 'Monica and Dr. Burke kiss'], what is the sequential order in this episode?


['Rachel apologizes to Ross and suggests a romantic dinner to make it up to him', 'Monica goes to her eye appointment with Dr. Burke', 'Monica and Dr. Burke kiss']



00:00 22:05 Time

Choose the correct option for the following question: Looking at these events ['Phoebe offers to waitress for Monica instead of Rachel', 'Monica and Phoebe arrive at Dr. Burke's apartment for a catering job', 'Dr. Burke tells Monica about his divorce', 'Chandler orders a pizza for him and Joey'], how do they unfold in the episode?

['Phoebe offers to waitress for Monica instead of Rachel', 'Monica and Phoebe arrive at Dr. Burke's apartment for a catering job', 'Chandler orders a pizza for him and Joey', 'Dr. Burke tells Monica about his divorce']

 Why the answer is not valid ?

Because Dr. Burke tells Monica about his divorce' before 'Chandler orders a pizza for him'

Figure 11: Examples of success and failure cases in Temporal Order of Events questions.

Deep context understanding:

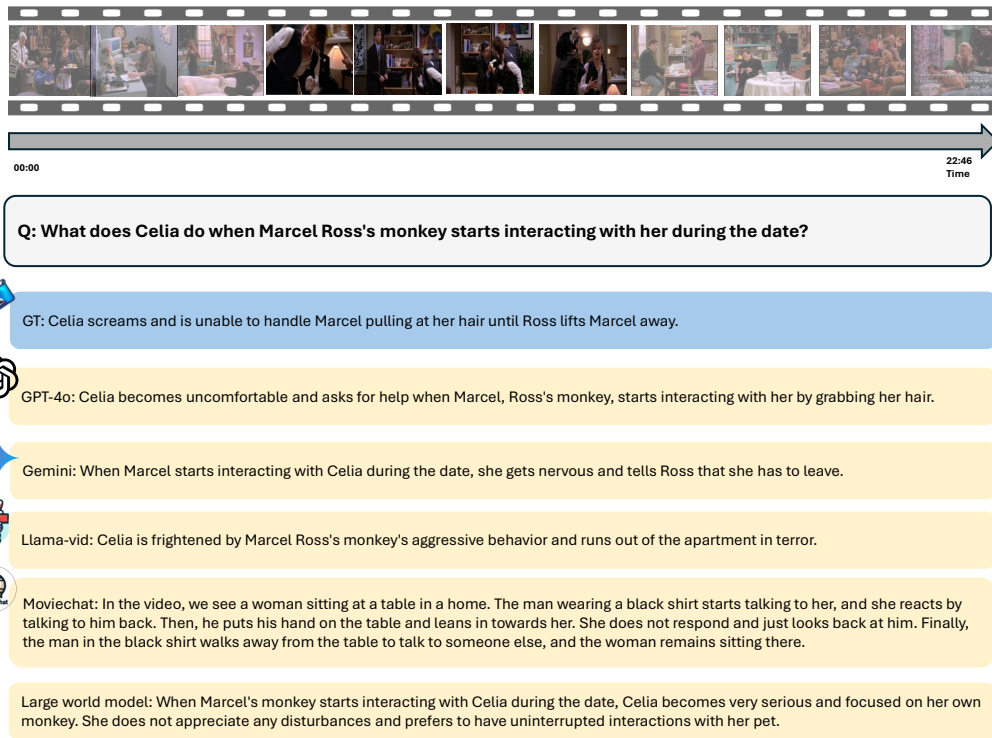



Figure 12: Qualitative results of deep context understanding questions

486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

Global appearance:



Q: Choose the correct option for the following question: Can you track the sequence of Penny's outfit changes in this episode?

GT: Option 1: red floral dress over a red shirt ,gray tank top over a pink sports bra , red and pink floral dress with a blue tank top underneath mustard yellow vest over a white blouse , blue tank top.

GPT-4o : option 0

Gemini : **Option 1:** ["red floral dress over a red shirt", "gray tank top over a pink sports bra", "red and pink floral dress with a blue tank top underneath", "mustard yellow vest over a white blouse", "blue tank top"]

Llama-vid : "0"


Moviechat: Next, a man and woman are seen talking while standing in the bathroom. The man is wearing glasses, and they are talking at the black mirror in the bathroom. A group of people are sitting in a red chair in front of the fireplace. The man is wearing glasses and is standing behind the white couch. A group of people are seen inside the house, including a woman wearing a purple dress. A man is seen in front of the television in the living room. Finally, the man and woman are sitting in a chair in front of the TV. The TV has a white case and a black frame. The man is wearing a pink shirt and the woman is wearing a blue dress.

Large world model : Yes, in the video, the man who is talking to Penny is seen putting on a black vest. Later, he is seen putting on a red and white shirt, and then he is seen putting on a red and white vest.

Figure 13: Qualitative results of global appearance skill


540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593


Scene transitions:





00:00 22:46 Time


Q :Choose the correct option for the following question: What is the chronological order of scenes in this episode?

 Option 4 : Monica and Rachel's apartment , Central Perk , Madison Square Garden , Ross's apartment , Madison Square Garden, Duncan's dressing room , Ross's apartment, Outside in the hallway , The street

 GPT-4o: option 0

 Gemini : ****Answer:**** option 5: ["Ross's apartment", "Ross's apartment, Outside in the hallway", "Madison Square Garden, Duncan's dressing room", "Monica and Rachel's apartment", "Madison Square Garden", "The street", "Central Perk"]

 Llama-vid : 1. Ross and Julie are in Ross' apartment.\n2. Ross and Julie are on their way to Madison Square Garden.\n3. Ross and Julie are in Duncan's dressing room at Madison Square Garden.\n4. Rachel and Monica are in their apartment.\n5. Rachel and Monica are at Central Perk.\n6. Rachel and Monica are in their apartment again.

 Moviechat: First, we see a group of friends gathering around a bar in a New York apartment. A young man is seen with his hand on a woman's leg. Next, the group is seen sitting around a table in the apartment. Then, a man and a woman are seen in a dark room with a green table and a dark room with a man in a white shirt. The group is gathered around the table in the first scene, and then the men and the woman are seen in the dark room. They are then joined by another man in the green room. Lastly, the group is seen in a dark room with a green table. There is a man and a woman sitting on a couch, and the woman is crying. The group is again seen in a green room with a couch, and a woman is sitting on it. They are also seen in a dark room with a green table and a man and a woman sitting in a green room. The group is seen in a green room with a couch, and a woman is crying. Finally, the group is seen in a dark room with a green table and a man and a woman sitting in a green room.

Large world model :Yes, the scene changes from Monica and Rachel sitting on the floor to a man standing in front of them. The man then proceeds to cut his hair with a knife, and Monica and Rachel start to panic. The man then uses a comb to brush his hair, and the three of them start to panic again.

Figure 14: Qualitative results of scenes transition skill

Spoiler questions



Q: Why didn't the Arquillian in the jeweler's head simply tell Jay that the galaxy was on his cat's collar?

To add a bit of mystery to the story. If he'd said, 'the galaxy in the jewel on the cat's collar', the movie would have ended much faster. Actually, Arquillian was indeed trying to tell Jay that the galaxy was on the cat's collar. He just didn't have the correct vocabulary to do so. Note how he stumbles over the word "war". He almost certainly thinks "belt" is the correct word for "collar", which is understandable because the articles of clothing are identical, as the only differences are that one is worn around the waist and the other is worn around the neck. And the cat's name is Orion, so he's being accurately descriptive, not deceitful. It's likely that the Arquillian didn't understand much English and that the Jeweler's body had a translator in it when conversing with humans. It was likely damaged when Edgar stabbed it through the neck.

GPT-4o: The Arquillian in the jeweler's head didn't simply tell Jay that the galaxy was on his cat's collar because he was using cryptic and indirect communication to convey the information, saying 'The galaxy is on Orion's belt.' This cryptic message is in place to protect the information from being easily understood, even under duress.

Gemini : The Arquillian in the jeweler's head didn't simply tell Jay that the galaxy was on his cat's collar because of a combination of factors:
Limited communication: The Arquillian's physical form was weak and dying. It might not have had the capacity to speak clearly.
Symbolic language: The Arquillian may have used a metaphorical or symbolic language that Jay didn't understand at that moment.
Time constraint: The Arquillian was likely very weak and dying. It may not have had enough time to explain the location of the Galaxy clearly.
It's crucial to remember that the movie relies on a mix of science fiction and humor. While the Arquillian's cryptic message adds to the mystery, it also adds an element of humor and absurdity to the film's plot.

Llama-vid : The Arquillian in the jeweler's head was too occupied with his own situation to reveal the galaxy's location. When Jay first approaches him, the Arquillian is trying to escape from the Arquillians who are chasing him and his ship. He is focused on his own survival and does not realize Jay is also a Men in Black agent. When Jay flashes his neuralizer, the Arquillian realizes Jay is not an alien and is not a threat. At this point, the Arquillian is just trying to get away from the Arquillians and their ship. He does not want to reveal any information that could help the Arquillians find him.

Moviechat : In this video, we see a man in a grey shirt walking through a dimly lit building at night. As he walks, we see other people walking by as well. In the background, we can see the man in the blue shirt with a silver watch on his wrist. He is walking down the hallway as well. The man in the grey shirt is wearing a black shirt with a white collar as he goes down the stairs. There is a black and white photo of a person holding a camera that is displayed on the wall behind him. We see another man in the dark suit in the room with the grey shirted man, and a man in the black and white shirt in the room as well. Finally, we see a man in a black and white outfit looking up and down the hallway.

Large world model : The Arquillian in the jeweler's head simply told Jay that the galaxy was on his cat's collar as a playful and humorous interaction. It seems to be a lighthearted and amusing moment between the two. The Arquillian in the jeweler's head was likely a character or a character in the jeweler's head who was present during the interaction. By simply telling Jay that the galaxy was on his cat's collar, the Arquillian was adding a touch of humor and entertainment to their conversation.

Figure 15: Qualitative results of spoiler questions skill

Linking multiple events:

System prompt :

You play two roles: a human asking questions related to a video and an intelligent chatbot designed to help people find information from a given video.

Your task is to generate question-answer pairs specifically related to linking multiple events in the video content.

You will first play the role of a human who asks questions that link multiple events together in the video, and then play the role of an AI assistant that provides information based on the video content.

##TASK:

Users will provide information about the video, and you will generate a conversation-like question-and-answer pairs specifically focusing on linking multiple events together in the video to make the questions comprehensive across the video.

Generate TWENTY descriptive and conversational-style questions and their detailed answers based on the given information, specifically related to linking multiple events together in the video.

##INSTRUCTIONS:

- The questions must be conversational, as if a human is asking them, and should directly relate to linking multiple events together in the video.

- The answers must be detailed, descriptive, and should directly reference the information provided.

- The number of events to link together can vary from 2 to any number of events.

Please generate the response in the form of a list of Python dictionaries as strings with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text respectively.

For example, your response should look like this: [{"Q": "Your question here...", "A": "Your answer here..."}, {"Q": "Your question here...", "A": "Your answer here..."}].

Make sure to avoid to put double quotes inside string with double quotes, use single quotes instead. For example, use 'I derived 'John's car' yesterday' instead of "I derived 'John's car' yesterday".

please only output the required format, do not include any additional information.

Remember well the output format of ONLY a PYTHON LIST as output and DON'T output the python shell because I will use python ast library to parse your output list.

Few shot examples about the questions:

- What is the influence of event A on event B?

- How does event A lead to event B?

- What is the relationship between event A and event B?

- What is the impact of event A on event B?

- What is the connection between event A, event B, and event C?

User prompt:

The user input is (summary).

Please generate the response in the form of a PYTHON LIST OF DICTIONARIES as strings with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text respectively.

For example, your response should look like this: [{"Q": "Your question here...", "A": "Your answer here..."}, {"Q": "Your question here...", "A": "Your answer here..."}].

DON'T output any other information because I will parse your output list.

Figure 16: Detailed prompt for Linking multiple events questions generation

Character actions:

System prompt:

You play two roles: a human asking questions related to a video and an intelligent chatbot designed to help people find information from a given video.

Your task is to generate a question-answer pairs specifically related to each character actions through the whole video content.

Your task is to first play the role of a human who asks questions about each character actions through the whole video content. and then play the role of an AI assistant that provides information based on the video content.

##TASK:

Users will provide information about a video, and you will generate a conversation-like question and answers pair specifically focusing on each character actions through the whole video content.

Generate one question for each character that summarize all the actions did through the whole video content.

##INSTRUCTIONS:

- The questions must be like a human conversation and directly related to each character actions through the whole video content.

- The answer must be detailed and descriptive that summarize all actions for each character in the video and should directly reference the information provided.

- Focus on both the visual and textual actions but focus more on the vision actions as these questions are designed for video understanding.

##SAMPLE QUESTIONS:

- ('Q1': 'What did ross do through this video?', 'A': 'At the beginning of the episode he drank coffee in central park , then went to his apartment then ate some pizza.')

- ('Q1': 'Summarize all actions that chandler did in this video.', 'A': 'At the beginning of the episode he read a magazine then went to his work by taxi , and finally he went to Monica's apartment to set with his friends.')

User prompt:

This is the episode summary: (caption). \n

This is the episode script: (script). \n

Please generate the response in the form of list of Python dictionaries string with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text, respectively.

For the answer, please make it as a python list of actions in chronological order

For example, your response should look like this: [['Q': 'Your question here...', 'A': ['Action 1','Action 2',...]],['Q': 'Your question here', 'A': ['Action 1','Action 2',...]]].

Please be very accurate and detailed in your response. Thank you!

Figure 17: Detailed prompt for sequence of character actions questions generation

Temporal order of events:

System prompt:

You play two roles: a human asking questions related to a video and an intelligent chatbot designed to help people find information from a given video.

##TASK:

Users will provide an episode Screenplay Script. Your task is to extract the events from this Screenplay Script. Ensure that the events are listed in chronological order

First read the Screenplay Script and think carefully to extract the all events.

##Few shot samples

Episode Screenplay Script: {user Screenplay Script}

Extract the events from this episode Screenplay Script:

The response should be in the format: ['Event A', 'Event B', 'Event C', 'Event D',...], ensuring that the event B is after event A and before Event C.

Remember well the output format of ONLY a PYTHON LIST of events and DON'T output the python shell because I will use python ast library to parse your output list.

User prompt:

Episode Screenplay Script: {script}

Extract the events from the Screenplay Script in a list

please provide the response in the format of PYTHON LIST of DON'T output any other information because I will parse your output list.

DON'T output any ' or ' in your response but use /u2019 for ' and /u2019s for 's and /u2019t for 't and s/u2019 for 's' or 's'

Figure 18: Detailed prompt for Temporal order of events questions generation

Scene transitions:

System prompt:

##TASK:

Users will provide an episode Screenplay Script. Your task is to extract scene transitions in from this script.

First read the Screenplay Script and think carefully to extract the transitions.

##Few shot samples

Episode Screenplay Script: {user Screenplay Script}

Extract the scene transitions from this episode Screenplay Script:

please provide the response in the format of PYTHON LIST of scene transitions like this example : ['scene A name', 'scene B name', 'scene C name',...], ensuring that the scene changed

from A to B then C and so on.

Scene names should be places name or location names where the scene is taking place such as home , cafe , bar , car and so on.

User prompt:

Episode Screenplay Script: {script}

Extract the scene transitions from this Screenplay Script in a list

please provide the response in the format of PYTHON LIST of scene transitions like this example : ['scene A name', 'scene B name', 'scene C name',...], ensuring that the scene changed

from A to B then C and so on.

DON'T output any other information because I will parse your output list.

Figure 19: Detailed prompt for scene transitions questions generation

702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

Deep context understanding:

System prompt:

You play two roles: a human asking questions related to a video and an intelligent chatbot designed to help people find information from a given video.

##TASK:

Your task is to first play the role of a human who asks questions related to deep context understanding in the video and then play the role of an AI assistant that provides information based on the video content.

Users will provide human video summary and the video script, and you will generate a conversation-like question and answers pair specifically focusing on measuring the viewer's context understanding.

##INSTRUCTIONS:

- The questions must be conversational, as if a human is asking them, and should directly relate to deep context understanding for the video content.

- The answers must be detailed, descriptive, and should directly reference the information provided.

- The number of questions should be up to 20 questions and answers.

- The questions should be tricky and hard to answer to measure the viewer's context understanding.

- The answers must be detailed, descriptive, and should directly reference the information provided.

- It will be good if most of the questions are related to the visual content of the video.

- Again, the questions should be very tricky and hard to answer to measure the viewer's context understanding.

Please generate the response in the form of a list of Python dictionaries as strings with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text respectively.

For example, your response should look like this: [{"Q": "Your question here...", "A": "Your answer here..."}, {"Q": "Your question here...", "A": "Your answer here..."}].

please only output the required format, do not include any additional information.

If you want to type 's' or 't' and so on, please use '\u2019s' and '\u2019t' for 't' and so on.

Test your output by using the python ast library to parse your output list.

Remember well the output format of ONLY a PYTHON LIST as output

User prompt:

video summary: {caption}.

video transcript: {script}.

Please generate up to 20 questions and their answers in the form of list of Python dictionaries string with keys 'Q' for question and 'A' for answer. Each corresponding value should be the question-and-answer text respectively.

For example, your response should look like this: [{"Q": "Your question here...", "A": "Your answer here..."}, {"Q": "Your question here...", "A": "Your answer here..."}].

Figure 20: Detailed prompt for deep context understanding questions generation

REFERENCES

- Google Gemini. Gemini technical report, 2024. URL <https://deepmind.google/technologies/gemini/flash/>.
- Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models, 2023.
- Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023.