Supplementary Materials for FunQA

Binzhu Xie[*⋆1], Sicheng Zhang[*2], Zitang Zhou[*1],
Bo Li[3], Yuanhan Zhang[3], Jack Hessel[4], Jingkang Yang[3], and Ziwei Liu[3✉]

[1] Beijing University of Posts and Telecommunications, Beijing, China
[2] Khalifa University, Abu Dhabi, UAE
[3] S-Lab, Nanyang Technological University, Singapore
[4] The Allen Institute for AI, WA, USA
https://github.com/Nicous20/FunQA

A More Details of FunQA Dataset

A.1 Dataset Construction Pipeline

Video Selection In constructing the dataset, we adhered to three principles to address the challenges in video understanding capabilities: our dataset, FunQA, is **visual centered** and emphasizes **counter-intuitive reasoning**, **spatial-temporal reasoning**. Based on these principles, we collect 4365 videos from 3 different art genres and created three subsets: HumorQA, CreativeQA, and MagicQA.

HumorQA HumorQA composed of 1,769 meticulously curated web videos, serves as a unique source of insight into human humor comprehension. Notably, it contains the shortest average video length of 7s among the three subsets. We believe that the human process of understanding humor is complex and deep, requiring a holistic understanding of the video and adding a degree of common sense to it. Psychological research has demonstrated that humor arises from the incongruity [8, 18] between reality and expectations, flourishing with the skillful juxtaposition and transformation of events [1,10,12]. This makes humorous videos a valuable asset for the VideoQA dataset, anticipated to enhance a model's proficiency in integrating information and performing deep reasoning.

CreativeQA CreativeQA is a collection of 927 videos averaging 48s in length from a TV show called Kasou Taishou [3]. This program, showcasing original and novel skits performed by various amateur groups and judged by a panel, boasts a strong creative flair [22]. The essence of the show lies in using a mix of people and props to mimic reality, with audiences deriving pleasure from information integration and comparison. We anticipate that the imitation nature of the show will challenge the model's capacity for information extraction, while the longer video length and need for understanding creativity will put to test the model's comprehension of spatial-temporal information.

MagicQA MagicQA encapsulates 1672 magic performance videos sourced from across the web, spanning various genres like camera magic, close-up magic, and

⋆ * indicates equal contribution. ✉ Corresponding author. Contact: ziwei.liu@ntu.edu.sg

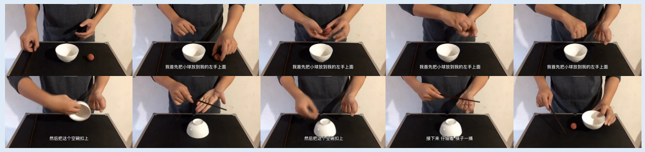
Table A1: Comparison between FunQA and other existing benchmarks (Complete Version). Compared to other datasets, FunQA revolves around the captivating realm of interesting and counter-intuitive videos. The tasks within FunQA are specifically designed to challenge the vision capabilities of models, requiring strong skills in producing an in-depth description, interpretation, and spatial-temporal reasoning. Here we clarify the abbreviation in the table. For annotation type: 🧑 denotes Manual Annotation and 🤖 for Automatic Annotation; **Avg Len** denotes video average length; # **Clips** means number of video clips; **VC** for visual-centric, **Des.** for Description, **Exp.** for Explanation, **STR** for Spatial-temporal Reasoning, **MC** means Multiple Choice QA, and **OE** shows Open Ended QA with **Average Word Count** per response.

Dataset	Domain	🧑 or 🤖	Video		Question Answer						
			Avg Len	# Clips	# QA	VC	Des.	Exp.	STR	MC	OE
TGIF-QA [7]	Social Media	🤖	3s	72K	165K	✓	✓	✗	✓	✗	2.1
MSRVTT-QA [27]	Social Media	🤖	15s	10K	244K	✓	✗	✗	✓	✗	1.0
ActivityNet-QA [34]	Social Media	🧑	180s	6K	58K	✓	✗	✗	✓	✗	1.9
MSVD-QA [27]	Social Media	🤖	10s	2K	51K	✓	✗	✗	✗	✗	1.0
YouTube2Text-QA [32]	Social Media	🤖	10s	10K	123K	✓	✗	✗	✓	✓	N/A
AGQA [6]	Social Media	🤖	30K	10s	192K	✓	✗	✓	✓	✗	TBD
AVQA [30]	Social Media	🧑	60s	9K	57K	✓	✗	✓	✓	✓	N/A
NExT-QA [26]	Daily life	🧑	44s	5K	52K	✓	✓	✓	✓	✓	2.6
Social-IQ [35]	Daily life	🧑	99s	1K	8K	✓	✗	✓	✗	✓	N/A
STAR [25]	Daily life	🤖	-	23K	60K	✓	✗	✗	✓	✓	N/A
FIBER [2]	Daily life	🧑	10s	28K	2K	✗	✓	✗	✓	✗	TBD
MovieQA [23]	TV shows	🤖	203s	7K	6K	✗	✗	✓	✓	✓	N/A
TVQA [13]	TV shows	🧑	76s	22K	153K	✗	✗	✓	✗	✓	N/A
TVQA+ [14]	TV shows	🧑	8s	4K	30K	✗	✗	✓	✓	✓	N/A
KnowIT-VQA [5]	TV shows	🧑	60s	12K	24K	✓	✗	✓	✓	✓	N/A
SUTD-TrafficQA [28]	Traffic	🧑	5s	10K	623K	✓	✗	✗	✓	✓	N/A
MarioQA [20]	Games	🧑	5s	188K	188K	✓	✗	✓	✓	✗	2.0
CLEVRER [33]	Synthetic Videos	🧑	5s	20K	305K	✓	✗	✓	✓	✓	N/A
Env-QA [4]	Egocentric	🧑	20s	23K	85K	✓	✗	✗	✓	✓	N/A
HumorQA (Ours)	Daily life	🧑	7s	2K	141K	✓	✓	✓	✓	✓	**28.2**
CreativeQA (Ours)	Performance	🧑	48s	1K	79K	✓	✓	✓	✓	✓	**59.1**
MagicQA (Ours)	Magic shows	🧑	10s	2K	92K	✓	✓	✓	✓	✓	**27.6**
FunQA (Ours)	**Surprising Videos**	🧑	19s	4K	**312K**	✓	✓	✓	✓	✓	**34.2**

stage magic. The essence of magic revolves around the creation of seemingly impossible illusions [21], employing diverse effects such as disappearance, creation, and transformation. These illusions are infused with abundant spatial-temporal information. Through this dataset, we aim to empower the model to not only track the ensuing changes in objects but also unravel the underlying mechanics [11] of these transformations.

Pre-processing & Qualification For videos related to humor and magic, we downloaded them from different streaming platforms, mostly in the form of compilations. For Creative videos, we downloaded 26 episodes publicly available from Tokyo TV in Japan. We provided rigorous training to the annotators to ensure high-quality video clips in the final compilation. Annotators who successfully completed the Clip task according to the requirements are considered qualified and can proceed to the next stage of annotation.

Training & Annotation We conducted systematic training for all annotators who passed the previous round of annotation, focusing on different tasks:
For the *Counter-intuitive Timestamp Localization Task (H1, C1, and M1)*, the annotation format is a pair of numbers enclosed in square brackets, [xxxx, xxxx]. We asked the annotators to record the time intervals in which they felt pleasure (or amusement or shock) while watching the video.
For the *Task Detailed Description Task (H2, C2, and M2)*, we requested objective



M4: What is the principle behind this magic trick?
Answer:
The man picks up the small ball with his right hand and puts it into his left hand and clutches it, this is a fake action, in fact the ball is still in the man's right hand, when he picks up the bowl with his right hand, he uses the cover of his fingers to put the ball into the bowl. The ball is in the bowl when the bowl is buckled on, and the man uses a thin stick behind him just to pretend to put the ball through the bowl on the surface of the bowl to produce the magic effect.

Fig. A1: Example of Magic Method Task. During the annotation process, we discovered that even as humans, it is difficult to fully understand the complete principles behind the implementation of magic tricks right from the beginning.

FunQA Video Selection & Clip Principle

1. The video content must be fun (any point that feels funny is acceptable: it can be a series of unfortunate events, jokes, etc.). If no funny moments are found, delete the segment.
2. Avoid sensitive topics such as racial discrimination ✗, gender opposition ✗, personal attacks ✗, national politics ✗, sexual orientation ✗, etc.
Examples: *Jokes about black people in a derogatory manner.* ✗
Gender wars, content related to disabled individuals. ✗
Implicit references to political leader. ✗
3. Avoid including extremely serious accident.
Examples: *Visible bleeding.* ✗
Bone fractures. ✗
Car accidents that clearly result in severe injuries or fatalities. ✗

Clip：Divide the original video into multiple segments.
Note: These intervals should be as continuous as possible and may overlap, but there should be no gaps in between. For Examples:
Correct：*[0000, 0012], [0012, 0022], [0022, 0034]*
Or：*[0000, 0012], [0011, 0022], [0020, 0034]*
Incorrect：*[0000, 0012], [0015, 0022], [0025, 0034]*

Fig. A2: FunQA Video Selection and Clip Principle. We have a zero-tolerance policy regarding the inclusion of offensive content in our dataset. During the video sourcing process (video selection and video clipping stages), we ensure that such content is completely eliminated.

descriptions of what happened at [xxxx, xxxx], emphasizing a "what you see is what you get" approach. It is important to note that the annotations should only cover the selected time intervals and should not include subjective adverbs (such as vividly, vividly, or wildly). When describing characters or objects, be concise and add modifiers if there is ambiguity. An example of a poor label is "a man and a woman," while a relatively better example is "a man wearing a red hat and a woman wearing an apron."

For the *Counter-intuitiveness Reasoning Task (H3, C3, and M3)*, explain why the video is interesting in the context of the overall content. This part requires interpretive answers based on analysis, reasoning, and prior knowledge to explain why the video is counter-intuitive.

Specifically for the *Magic Method Task (M4)*, we found that most annotators were not professional magicians, and even when watching instructional magic

Annotation Content	Task	Task requirement	Reason for Incorrection	Modifications
Short description of H2.				
Three Chickens Crawling	H2	Description of the chosen moment	The content is too brief. The description in H3 is clear enough and should be placed entirely in H2.	In the farm, there are three chickens: two black ones on each side and one white one in the middle, crawling synchronously.
H2 has subjective coloration.				
The man attempted to jump over the garbage bin but ended up getting hit in the groin by the bin.	H2	Description of the chosen moment	"Attempted to skip" is not an objective description. Speculations about this aspect of psychological activity should be written in H3.	A man faced the trash bin, with both hands propped on it, and his legs lifted off the ground as he swung forward, then he sat on the trash bin with a thud and flew out in the end.
H2 has subjective coloration.				
The man prepared to jump over the fence, exerted force to leap, but failed to jump high enough and ultimately stumbled and fell.	H2	Description of the chosen moment	"Parpared to jumo" is not an objective description. Speculation about this aspect of psychological activity should be written in H3.	A man leaped towards the fence, but his leg got tripped, causing him to fall straight down and break the fence as well.
H2 has subjective coloration.				
The adult was initially supposed to carry the child onto the electric bike, but the adult rode away before the child could get on. The child chased after them from behind.	H2	Description of the chosen moment	"The adult was initially supposed to carry the child" is not an objective description. The explanation for this behavior should be written in H3."	A child stands behind the car, while the adult drives away. The child runs after the electric bike.
Loss of important information such as sound or narration.				
A scene where a man does push-ups while hitting a metal bowl with his head to imitate the sound of a telegraph being clicked.	C3	Why creative	During the performance, the actor says, "SOS SOS, a ship is in distress in Tokyo Bay." This information is crucial to help the audience understand that they are imitating a distress telegraph message. The role of audio cues should be incorporated into the answer to emphasize its importance.	A man does push-ups while hitting a metal bowl with his head and simultaneously recites lines like 'SOS, SOS,' imitating the scene of a telegraph machine working when sending a distress signal.
Missing content in the explanation of the reasons.				
The small ball in the palm of the hand cannot disappear into thin air, so it should not appear inside the cup.	M3	Why magic	Actually, there are two effects involved: the disappearance of the ball from the hand and the appearance of the ball inside the cup. Therefore, the logical word "so" is not appropriate. The logic should be modified, and all the effects and their reasons for being unreasonable should be fully described.	Objects cannot vanish or appear out of thin air, so the ball in the hand cannot vanish inexplicably, nor can the ball appear inside the cup without a logical explanation.

Fig. A3: Examples of incorrect annotations and the suggested modifications. Here we clarify the meanings of different cell colors. **Red**: Errors made by the annotator during the annotation process. **Yellow**: Types of annotation errors. **Green**: Reasons for the errors and suggested corrections.

videos, it was challenging for them to provide complete and reasonable annotations for magic methods. Therefore, for M4, we only provided a small number of examples, which do not exist in the FunQA dataset. Fig. A1 shows an example of M4.

We conducted strict quality reviews in real-time for the annotations, including but not limited to checking if the annotations meet the requirements, if the selected videos contain sensitive content, and if the annotation texts contain sensitive content. The annotation principles (video selection and video clip, etc) can be seen in Fig. A2. Fig. A3 also illustrates some errors encountered during the annotation process and the suggested modifications provided. We assured that every annotation included in the final release of FunQA has been subjected to rigorous **multi-person, multi-round** review processes. Specifically, every piece of released FunQA version annotated content underwent scrutiny by three distinct annotators. After completing all the annotations, we conducted a consensus evaluation study to assess the objectivity of our annotations. We randomly selected 10% of the videos and asked annotators to provide a consistency score (high consensus, medium consensus, low consensus) between their first and second annotations, considering all the previous annotations shown to them.

Copyright & License We respect the copyright of each video. Our data sharing approach is informed by the models established by previous studie like Kinetics [9], HD-VILA-100M [29], and others. Instead of providing the original raw data, we only supply the YouTube video IDs necessary for downloading the respective content.

GPT Prompt for FunQA Translation and Expansion

Now I have two tasks for you to complete at once.
Task 1. CH2Eng Translation: Please **translate** the content into English.
Task 2. Expansion_with_Eng: Please generate five **synonyms** according to the content just translated into English.

User Input

一只白色的猫咪整个趴在厕所蹲厕里。

GPT-3.5 Output

CH2Eng Translation: A white cat is lying flat inside the toilet bowl.
Expansion_with_Eng_1: A white cat resting entirely in a toilet bowl.
Expansion_with_Eng_2: The entire body of a white cat is sprawled across the toilet seat.
Expansion_with_Eng_3: The white cat is The white cat is lying flat within the toilet's squatting position.
Expansion_with_Eng_4: A white-coloured feline is stretched out inside the toilet bowl.
Expansion_with_Eng_5: The complete posture of a white cat can be seen in the toilet, lying down.

Fig. A4: GPT prompt for Translation and augmentation. Under our carefully designed prompt, GPT-3.5 can automatically generate English translations and five synonymous sentences from our high-quality Chinese annotations, thereby expanding FunQA dataset.

We respect the personal identity information of everyone appearing in our videos and always strive to eliminate offensive content. By conducting strict reviews of the annotators and real-time quality checks during each annotation process, we ensure the absence of offensive content.

A.2 Language Augmentation

For each QA pair, we first asked GPT-3.5 to translate the previous Chinese answers into English, and then let GPT-3.5 give five separate answers with the same meaning but different linguistic expressions. Afterward, we filtered out incorrect and incomplete generations. Fig. A4 shows the prompt we gave to GPT-3.5.

However, as mentioned in the main text in the limitation, the translation of GPT-3.5 will be ambiguous when the original text uses Chinese harmonics and some special Chinese meanings, and this problem mostly occurs in the Title Task, we screened such problems by hand and modified them, as shown in Fig. A5.

A.3 Data Statistics

For a complete comparison between FunQA and other existing benchmarks, please see Table A1. We collected raw videos from multiple video platforms, including short videos, long videos, and video clips, and the detailed data can be seen in Table A2. The statics of FunQA dataset before GPT-3.5 extension can be seen in Table A3. The word cloud of the all annotation word is shown in Fig. A6 (a-d).

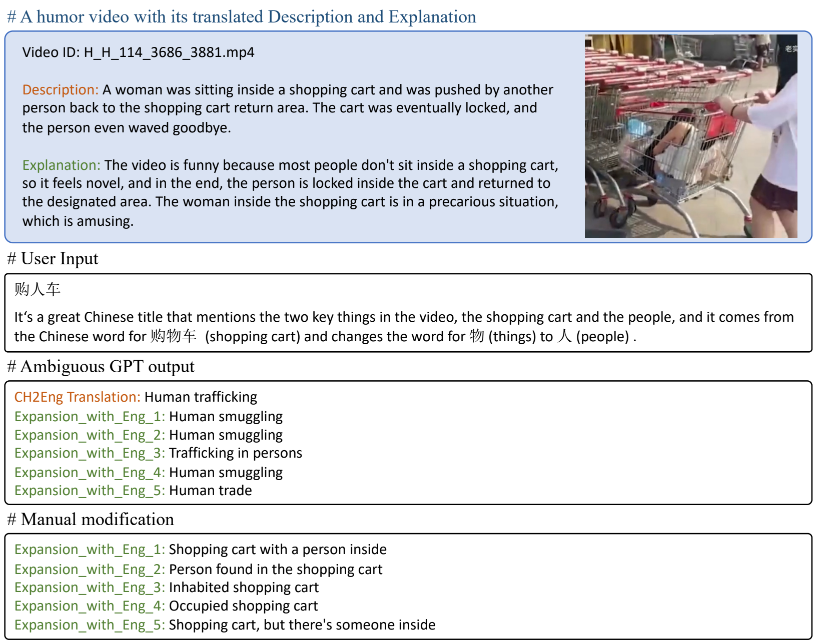


Fig. A5: Issues in translation caused by Chinese-English cultural differences. In this example of translating and expanding a humorous video title, GPT-3.5 failed to understand the meaning of the original Chinese title, and we filtered out such data and made manual changes.

Table A2: Statistics of the FunQA raw data.

FunQA Subset	Type	Source	# Videos	Avg.len (s)	Total.len (h)
Humor	Daily Life (Human)	Youtube	351	182	15.8
			1296	14	5.18
	Nature (Animal)		230	133	8.52
Creative	Performance	Youtube	26	6060	43.77
MagicQA	Close up Magic	Youtube	765	96	20.40
	Camera Magic		334	152	14.17
FunQA	-	-	3002	129	107.87

B FunQA Extension Datasets

Our main objective in designing these extended datasets is to leverage our high-quality annotated data and provide a rich and suitable data format for models. In addition, we also wanted to test the capability of GPT-3.5 and the quality of the dataset, and it turns out that GPT-3.5 can expand our data tens of times

Table A3: Statistics of the FunQA before GPT-3.5 extension.

Datasets	Avg length (s)	# Clips (K)	# QA pairs (K)	# QA per clip
HumorQA	7	1.8	7.2	4.0
CreativeQA	48	0.9	4.5	5.0
MagicQA	10	1.6	4.8	3.0
FunQA	19	4.3	16.5	3.8

(a) HumorQA. **(b) CreativeQA.** **(c) MagicQA.** **(d) FunQA.**

Fig. A6: Word cloud of FunQA free-text answer.

after providing high-quality free-text annotations, and our FunQA dataset is extremely scalable.

B.1 FunQA-MC Dataset

FunQA-MC (Multi-choice) Dataset is prepared to provide training and testing for arbitrary models, in this dataset our QA pairs are in the form of multiple choice, the answer is a word, phrase, or short sentence, and the type of questions are all descriptions. An example of the prompt given to GPT-3.5 and the data generated can be seen in Fig. A7. For the **FunQA-MC-R**, which is a multi-choice version specifically containing counter-intuitive reasoning questions, we randomly chose 100 QA pairs from Task 2 (Counter-intuitiveness Reasoning Task) to create the subset.

B.2 FunQA-DIA Dataset

Most of the current LLMs are in the form of dialogues. To cater to their data input, we produced the FunQA-DIA (Dialog) dataset, in which we used GPT-3.5 to convert QA pairs into recursive dialogues with added context. In addition to this, we also let GPT-3.5 freely generate future dialogue development, which greatly expands the amount of information. An example of the prompt given to GPT-3.5 and the data generated can be seen in Fig. A8.

GPT Prompt for FunQA Multi-choice

You are now a data augmentation assistant. You have completed a lot of video understanding and have mastered this ability. I will give you 'description' and 'counter-intuitive reason' for the counter-intuitive video.
You will help me to complete the understanding of this paragraph and generate **5 English QA pairs** from it.
The form is multiple choice. Each question has **5 choices**, one of which is the correct answer and the other four are interference items.
You need to mark the correct answer. The question can only be **what, how, which or who**. The answer is a **word, phrase, or sentence** of 15 words or less.

User Input

description: A white cat is lying flat inside the toilet bowl.
counter-intuitive reason: The cat is so comfortable lying on its stomach in the toilet that it is both helpless and funny.

GPT-3.5 Output

Q: What is the main subject of the video?
1. A black cat
2. A white cat
3. A dog
4. A bird
5. A fish

Q: How is the white cat positioned inside the toilet bowl?
1. Standing on its hind legs
2. Sleeping on its side
3. Lying flat on its stomach
4. Sitting upright
5. Hanging upside down

Q: Which word best describes the cat's behavior in the video?
1. Annoyed
2. Excited
3. Curious
4. Comfortable
5. Scared

Q: What is the cat's posture inside the toilet bowl?
1. Upside down
2. Leaning against the edge
3. Sitting like a human
4. Balanced on one leg
5. Lying flat on its stomach

Q: Who or what finds the cat's position in the toilet bowl funny?
1. The cat itself
2. A person watching the video
3. Another animal in the room
4. Nobody finds it funny
5. The cat's owner

Fig. A7: GPT prompt for FunQA Multi-choice. We give GPT-3.5 prompt as system, and then enter the description and explanation of the video, and GPT-3.5 will output the data in the form of multiple-choice questions according to the prompt.

GPT Prompt for FunQA Dialog

You are now my dialogue generation assistant, you have done a lot of video understanding and mastered this ability, I will give you a 'description' of the counter-intuitive video and a 'counter-intuitive reason' of why it is counter-intuitive.
Where 'description' is an objective description of a video, including what happens in the video and 'counter-intuitive reason' is an explanation of why the video is counter-intuitive.
You're going to use that information to generate **a question and answer** for me, using the text that I give you, start with descriptive questions like **what, when, which, who, where**, and then start with **why, how**, and then you're going to give some **open questions**.
That is to say, the question is gradual, and you should try to use the previous information later. All of the above answers are short sentences, and try not to exceed 30 words

User Input

description: The barber gave the woman a haircut. After the woman cut her hair, her hair stood on end. Next to her appeared a parrot with the same hairstyle as the woman.
counter-intuitive reason : Women look funny when they get their hair cut and look like parrot hair.

GPT-3.5 Output

Human: What is happening in the video?
Model: A girl is getting her hair cut by a hairdresser. After the haircut, her hair stands up and a parrot appears next to her with the same hairstyle.
Human: Why is the video humorous?
Model: The girl's hair standing up and matching the parrot's hairstyle is unexpected and funny.
Human: What this girl's mood might be like?
Model: She may be dissatisfied because generally speaking, a human's hair does not look like a parrot, which is not considered aesthetically pleasing.
Human: Does this parrot feel humor because the girl's hair?
Model: Probably not, it does not feel this, probably because its intelligence can not understand the relationship between the girl's hair and its hair.
Human: Do you think the parrot intentionally matched the girl's hairstyle?
Model: It's unlikely that the parrot intentionally matched the girl's hairstyle. It's more likely a coincidence or the result of the parrot's natural feathers.
Human: What do you think the girl will do afterwards?
Model: It's unlikely that the parrot intentionally matched the girl's hairstyle. It's more likely a coincidence or the result of the parrot's natural feathers.

Fig. A8: GPT prompt for FunQA Dialog. We give GPT prompt as system, and then input the description and explanation of the video, and GPT-3.5 will output the data in the form of dialogue according to the prompt.

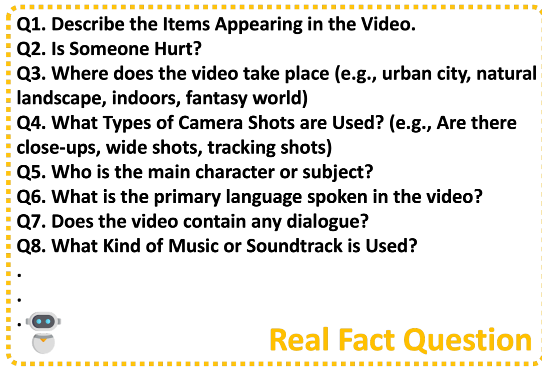


Fig. A9: Example of Real fact question by FunMentor.

C More Details of FunMentor

C.1 Real Fact Question

During the **Real Fact Collection** process, FunMentor relies on designing questions that VLMs can answer accurately and objectively to gain knowledge of the video's basic information. The specific questions are shown in Fig. A9.

C.2 Prompts Design

Based on the QA pairs collected as mentioned, FunMentor engages in multi-round dialogues with VLMs through our predefined prompts. This process involves providing increasingly precise instructions, guiding the VLM towards correctly answering the question. Specific examples of this are shown in Fig. A10.

D More Details of Experiment

The complete result of FunQA Benchmark (with traditional metircs' scores) are shown in Table A4.

D.1 Significance of New GPT-4 Based Metric

GPT-4 prompt design For each of the three tasks, we designed three prompts for scoring. *For Detailed Description Task (H2, C3, and M2)*, we designed the prompt in five areas: text length, text content variation, text detail variation, logical text description, and linguistic ability. For *Counter-intuitiveness Reasoning Task (H3, C3, and M3)*, we designed the prompt in six areas: expressiveness of language, the logic of response, the common sense of response, understanding of counter-intuition, differences in text detail, and length of text. For *Title Task (H4 and C4)*, we used the description, comprehension, and title of the manually

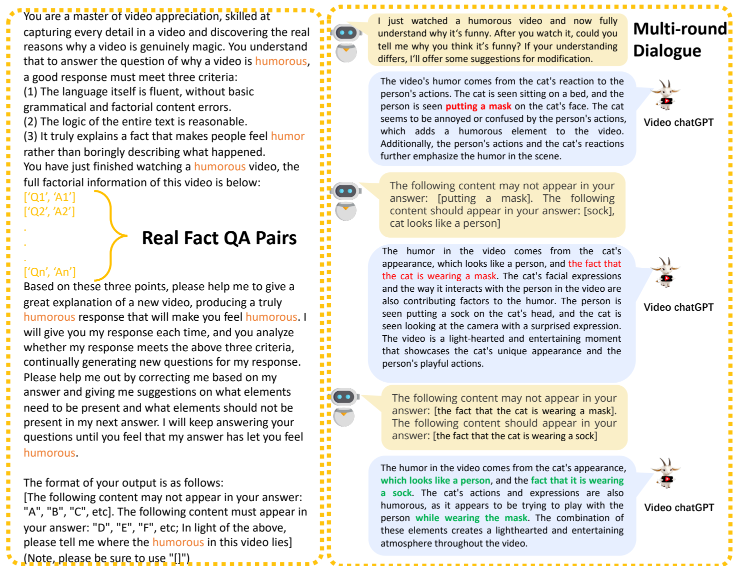


Fig. A10: Prompt Design of FunMentor in multi-round dialogue. On the left are the prompts from FunMentor, which incorporate the information from collected **Real Facts** and conduct **Answer Judgement** to **generate suggestions**.

annotated video as a reference to score the new title. Each Prompt can be seen in Fig. A11, A12, and A13.

Comparison between GPT-4 and Traditional Metrics The principle of the traditional metrics is relatively simple. The traditional metric principles are as follows.

BLEU-4 BLEU, full name is Bilingual Evaluation Understudy, is a commonly used machine translation evaluation metric. It evaluates how good a machine translation result is by comparing how well it matches the N-gram of one or more human-translated reference results, which is a sequence of N consecutive words. BLEU-4, i.e., evaluates how well a combination of two words (i.e., a binary) matches. BLEU introduces a correction factor, Brevity Penalty (BP), to avoid this problem, which penalizes machine translation if the result is shorter than the reference translation. The BLEU score is the geometric mean of the individual N-gram accuracy multiplied by the shortness penalty. That is, the score of BLEU takes into account the precision and length of the translation result.

ROUGE-L ROUGE, known as Recall-Oriented Understudy for Gisting Evaluation, is a commonly used evaluation method for tasks such as automatic digesting and machine translation. ROUGE is mainly evaluated by comparing the overlap between the generated abstracts and the reference abstracts. Among them, ROUGE-L is an important variant of ROUGE, where L stands for Longest Common Subsequence (LCS), i.e., the longest common subsequence. Unlike the n-gram, the longest common subsequence does not require consecutive occurrences of items in the sequence.

Table A4: Main Results on FunQA Benchmark (Complete Version). H2, C2, M2 represent the detailed video description task, and H3, C3, M3 represent reasoning around counter-intuitiveness. For the higher-level tasks, H4, C4 involve attributing a fitting and vivid title. The responses for all these tasks in free-text format. We use the following metrics: **BLEU-4 / ROUGE-L / CIDEr** (shown in the first row) and **BLEURT / GPT-4** (shown in the second row) for evaluation. C5 represents scoring the video creativity, and the metric is the **Accuracy** between the predicted score and the official score. We tested the caption-based and instruction-based models. Here we clarify the abbreviation in the table. **L.M.**: GIT_LARGE_MSRVTT; **L.V.**: GIT_LARGE_VATEX; **D.C.** means finetuned on Dense Caption; **FunQA** means finetuned on FunQA.

	HumorQA			CreativeQA				MagicQA	
Task	H2-Des.	H3-Rea.	H4-Title	C2-Des.	C3-Rea.	C4-Title	C5-Score	M2-Des.	M3-Rea.
- Caption-based Model									
mPLUG [16]	**1.5** / 16.4 / **1.0**	1.1 / 12.5 / 0.4	0.6 / 7.5 / 0.1	0.4 / 13.4 / 0.0	0.7 / 12.6 / 0.1	0.3 / 3.2 / 0.0		1.2 / 15.8 / 0.5	0.9 / 8.9 / 0.4
	19.9 / 3.9	**25.7** / **6.0**	22.1 / **11.2**	14.9 / 3.0	**24.2** / **6.9**	**20.8** / **18.8**	0.0	19.7 / 4.0	**21.2** / **8.1**
GIT (L.M.) [24]	0.5 / 12.8 / 0.2	0.0 / 0.0 / 0.0	**1.1** / 7.7 / **0.7**	0.0 / 6.40 / 0.0	0.0 / 0.0 / 0.0	0.3 / 1.5 / 0.2		0.2 / 11.2 / 0.1	0.0 / 0.0 / 0.0
	22.4 / 3.6	0.0 / 0.0	17.0 / 8.9	14.4 / 3.8	0.0 / 0.0	7.1 / 11.3	0.0	19.4 / 8.2	0.0 / 0.0
GIT (L.V.) [24]	1.2 / 16.9 / 0.6	0.0 / 0.0 / 0.0	1.0 / **8.8** / **0.7**	0.1 / 8.30 / 0.0	0.0 / 0.0 / 0.0	**0.5** / 2.8 / **0.4**		0.6 / 13.7 / 0.1	0.0 / 0.0 / 0.0
	33.3 / **4.0**	0.0 / 0.0	**25.9** / 10.0	**20.5** / **4.2**	0.0 / 0.0	10.5 / 12.0	0.0	**29.8** / **8.6**	0.0 / 0.0
- Instruction-based Model									
VideoChat [17]	0.5 / 13.7 / 0.0	0.5 / 13.5 / 0.0	0.8 / 5.1 / 0.5	0.3 / 7.50 / 0.0	0.3 / 7.70 / 0.0	0.2 / 1.2 / 0.2	67.5	0.6 / 15.5 / 0.0	0.3 / 9.2 / 0.0
	44.0 / 17.9	45.4 / 31.9	20.2 / 31.7	21.7 / 5.9	22.8 / **17.7**	7.3 / 31.1		47.4 / 8.2	43.1 / 44.6
Video-ChatGPT [19]	0.5 / 14.0 / 0.1	0.7 / 12.4 / 0.1	0.4 / 3.2 / 0.2	1.1 / **19.8** / 0.2	0.8 / 17.3 / 0.1	0.2 / 1.9 / 0.2	**85.4**	0.7 / 20.8 / 0.0	0.5 / 11.3 / 0.0
	39.9 / **24.3**	40.1 / 24.9	36.5 / 41.2	**45.8** / 6.6	45.2 / 9.1	30.9 / **48.8**		**50.8** / **11.2**	43.3 / 40.4
mPLUG-Owl [31]	0.5 / 13.3 / 0.2	0.5 / 13.2 / 0.2	0.6 / 5.3 / 0.0	0.8 / 16.5 / 0.1	0.7 / 13.3 / 0.0	0.1 / 1.1 / 0.1	66.7	0.4 / 16.7 / 0.0	0.5 / 8.4 / 0.0
	44.5 / 10,7	**47.3** / **35.0**	29.8 / **48.8**	43.0 / 5.0	**44.7** / 10.6	23.9 / 36.3		46.4 / 8.6	43.9 / 30.9
Video-LLaMA [36]	0.3 / 12.1 / 0.1	0.5 / 12.6 / 0.0	0.5 / 0.4 / 0.0	0.7 / 8.80 / 0.1	0.4 / 9.70 / 0.0	0.3 / 0.8 / 0.1	64.2	0.4 / 14.3 / 0.0	0.5 / 7.9 / 0.0
	48.4 / 7.7	42.9 / 29.0	46.5 / 34.1	45.5 / 7.2	41.1 / 17.2	42.3 / 31.2		50.1 / 10.2	39.0 / 28.0
Otter (D.C.) [15]	1.1 / 14.3 / 0.4	1.2 / 14.2 / 0.4	0.5 / 5.4 / 0.1	0.5 / 13.8 / 0.1	1.0 / 16.8 / 0.2	0.3 / 2.3 / 0.1	45.0	1.0 / 15.0 / 0.3	1.1 / 12.8 / 0.2
	30.2 / 7.7	32.3 / 28.3	21.7 / 20.0	28.7 / 1.7	32.9 / 7.9	17.7 / 36.3		32.5 / 2.1	27.3 / 36.8
Otter (FunQA) [15]	**1.5** / **18.1** / 0.9	**1.3** / **15.4** / **0.5**	0.8 / 5.9 / 0.5	**1.5** / 19.6 / **0.5**	**2.2** / **21.2** / **0.5**	0.3 / **4.3** / 0.3	69.4	**2.6** / **23.8** / **1.6**	**3.4** / **20.3** / **2.6**
	38.4 / 8.9	42.6 / 31.7	**47.5** / 32.1	40.0 / **7.3**	41.1 / 8.8	**44.5** / 38.8		44.7 / 10.3	**44.5** / **47.5**
Video-ChatGPT [19] + **FunMentor (Ours)**	0.6 / 14.2 / 0.6	1.1 / 14.4 / 0.3	1.0 / 5.7 / 0.6	1.4 / 20.2 / 0.3	1.4 / 18.8 / 0.2	0.3 / 2.2 / **0.3**	**85.4**	1.1 / 23.2 / 0.4	2.1 / 14.4 / 0.9
	65.2 / **33.2**	**57.5** / 36.5	50.2 / **65.1**	**66.3** / **14.2**	**58.7** / **23.4**	45.3 / **52.2**		**55.1** / **13.3**	**46.3** / 54.8
Otter (FunQA) [15] + **FunMentor (Ours)**	**1.9** / **20.2** / **1.1**	**1.7** / **19.3** / **0.9**	**0.9** / **6.8** / **0.9**	**2.4** / **23.1** / **1.1**	**3.3** / **26.5** / **0.7**	**0.4** / **4.7** / **0.3**	69.4	**3.1** / **25.5** / **2.5**	**5.6** / **21.2** / **3.4**
	33.4 / 13.4	37.8 / **45.8**	**58.3** / 34.2	60.4 / 11.0	44.4 / 9.3	**53.9** / 43.5		43.5 / 12.81	38.91 / **56.4**

CIDEr CIDEr, known as Consensus-based Image Description Evaluation, is an evaluation metric specifically designed for evaluating image description (Image Captioning) tasks. The main advantage of CIDEr is that it can capture more detailed information because it uses TF-IDF weights to emphasize n-grams that occur frequently in manual annotation but are not common in all image descriptions.

BLEURT BLEURT, full name Bilingual Evaluation Understudy with Representations from Transformers, is an evaluation method based on the transformer model, specifically for evaluating the output quality of machine translation and natural language generation tasks. Unlike traditional evaluation metrics such as BLEU and ROUGE, BLEURT does not directly compare the n-gram match between generated text and reference text but uses pre-trained language models (e.g., BERT) to understand the semantic information of text. the advantage of BLEURT is that it can capture the deep semantic information of text, and can address some problems that are difficult to be handled by traditional evaluation metrics (e.g. synonym substitution, utterance rearrangement) giving reasonable evaluation.

WUPS The Wu-Palmer similarity (WPUS) measure calculates how similar two word senses are. It considers the depths of the two synsets in the WordNet taxonomies, along with the depth of the Least Common Subsumer (LCS). The

GPT Prompt for Tile Evaluation

I will provide you with four text passages, in the following format [Description][Explanation][text1][text2], where the first two texts are the description of a video and the counter-intuitive explanation for it, and **the third text is a reference title**. By analyzing these two titles, you need to **judge whether the fourth text is a good title**. Please note that the fourth text might not be a title but a sentence containing the title of the video, like "the title of this video should be 'dog eating food'". In this case, you should extract "dog eating food" as the title, and then grade it. The **scoring range is 0-100**, and you need to consider the following points:
1. The title needs to mention the content of the video.
2. It is better if the title has a certain degree of humor or creativity.
Please give a score.

User Input

[The hairdresser blows the customer's hair with a hairdryer, causing all of the customer's hair to be blown away.]
[It's incredibly funny when the hairdresser uses the hairdryer on the customer's hair and ends up blowing it all away, leaving behind a completely bald head. The hairdresser's expression of shock adds to the humor.]
[Bald haircut]
[Hairdresser and patron in barbershop]

GPT-4 Output

Based on the description and explanation provided, the fourth title, "Hairdresser and patron in barbershop," is not very reflective of the video's content nor does it contain any humor or creativity. It simply describes a typical scene that could occur in any barbershop and does not give any indication of the unique and humorous event that occurred.
The third title, "Bald haircut," on the other hand, is a better title. It conveys the main outcome of the video (the customer becoming bald) in a succinct and slightly humorous manner.
Title Extracted: "Hairdresser and patron in barbershop"
Score: 20/100.
The score is relatively low due to the lack of specificity and humor. It might have been higher if the title incorporated elements of surprise, humor, or unexpected outcome, such as the customer's hair being completely blown away.

Fig. A11: GPT-4 prompt for Title Evaluation. We give GPT prompt as system, and then input a description, explanation, and two titles, the first one is our annotation, and the second one is the output of the model, and GPT-4 will evaluate the similarity between these two texts according to the prompt's requirements.

WUPS metric has certain limitations that make it difficult to use in VQA tasks. First, certain words are very similar in vocabulary, but their meanings can be very different. The problem may arise with color. For example, if the answer to a certain question is white and the system predicts the answer to be black, that answer will still get a WUPS score of 0.92, which seems high. Another limitation is that WUPS cannot be used for answers to phrases or sentences because it always deals with rigid semantic concepts, which are most likely to be single words. More examples as shown in A14.

It can be seen that the principles of the above traditional NLG metric are relatively simple and lack the ability to understand and evaluate the text with complex logic. In contrast, GPT-4, as an LLM, has a high ability to understand the text and can understand the text first and evaluate the similarity through prompt design.

Instability of GPT-4 responses During the experimental process using GPT-4 as the metric, we found that in a few cases, GPT-4 would provide different evaluations for the same content. In such cases, we would perform multiple evaluations and take the average. Fig. A15 shows an example of GPT-4 generating different responses to the same content.

GPT Prompt for Description Evaluation

Below, I will provide you with two text passages in the following format: [text1] [text2]. These two passages will serve as descriptions for a counter-intuitive (humorous, creative, magical) video. You are required to **evaluate text2** based on the following criteria:
1. Content: The content is worth a total of 20 points. If the content is almost identical, it scores 20 points. If the content has minor differences, it scores 15 points. If the content has significant differences, it scores 10 points. If the content has very major differences, it scores 5 points.
2. Details: Describe the details of the video clearly, including characters, settings, actions, dialogues, etc. This category is worth a total of 50 points. Deduct 5 points for each differing detail. Please note that you should explicitly identify and count the differing details when determining the score.
3. Coherence: The description should be logically consistent and should not include any illogical situations. Coherence is worth a total of 20 points. It is divided into three levels: If the coherence is almost identical, it scores 20 points. If the coherence is generally consistent but differs in details, it scores 15 points. If there are some differences in coherence but the overall logic is similar, it scores 10 points. If there are significant differences in coherence, it scores 5 points.
4. Language Expression (fluency, word choice, etc.): Language expression is worth a total of 10 points. It is divided into three levels: If the language expression is at a similar level, it scores 10 points. If there are minor differences in language expression, it scores 5 points. If there are significant differences in language expression, it scores 0 points.
5. Please note that if the content has **significant differences**, the total score should be multiplied by 0.5, and if the content has **very major differences**, the total score should be multiplied by 0.25.
In the end, you should provide a total score based on these five criteria.

User Input

[The person standing nearby struggles to open the bottle of tomato sauce and, in a moment of frustration, they inadvertently fling the bottle, causing the tomato sauce to spill all over a man's body and face.]
[The video depicts two women eating potato chips and fries while sitting on a bench, with a man standing behind them and staring at them.]

GPT-4 Output

Evaluation:
1. **Content:** The content has very major differences. The first text is about a person struggling with a bottle of tomato sauce, which eventually spills all over a man, while the second text is about two women eating snacks on a bench, with a man observing them. Score: 5/20.
2. **Details:** The details are entirely different. The characters' actions and interactions, the settings, and even the objects involved differ. A struggling person with a tomato sauce bottle has been replaced with two women eating snacks; an accidental spilling event is replaced by a man staring; the sauce bottle is replaced with potato chips and fries. The score deducted is 5 points for each differing detail, which amounts to 20 points (4 details). Score: 30/50.
3. **Coherence:** While each description independently makes sense and is coherent within itself, there's a stark contrast in terms of the narratives presented in the two texts. Score: 5/20.
4. **Language Expression:** The language expression differs significantly. The first text describes a comedic, albeit unfortunate, event, while the second text simply describes a somewhat normal scene. Score: 0/10.

Before applying the penalty for very major differences in content, the raw score is 5 (content) + 30 (details) + 5 (coherence) + 0 (language expression) = 40/100.
As the content has very major differences, we need to multiply the total score by 0.25. So, 40 * 0.25 = 10.
Final Total Score: 10/100.

Fig. A12: GPT-4 prompt for Description Evaluation. We give GPT-4 prompt as system, and then input two descriptions, the first one is our annotation, and the second one is the output of the model, and GPT-4 will evaluate the similarity between these two texts according to the prompt's requirements.

GPT Prompt for Explanation Evaluation

Below are two texts in the following format: [text1][text2]. These two texts will provide explanations for a counterintuitive video (humorous, creative, or magical). **For text2**, you need to provide a score based on the following criteria:
1. Language Expression (Fluency, Vocabulary): Total score of 5 points. It is divided into three levels. If the language expression of this text is at a level comparable to a score of 5, give 5 points. If there is a significant difference in language expression, give 0 points.
2. Logical Coherence: The explanation should demonstrate logical coherence, ideally using logical words and causal relationships. Total score of 10 points. It is divided into three levels. If the logical coherence is essentially the same as the reference answer, give 10 points. If there are minor differences in details but overall logical consistency, give 5 points. If there are significant differences in logic, but the overall logic is similar, give 5 points. If there are major differences in logic, give 0 points.
3. Common Sense Errors: Total score of 10 points. This text should not contain any obvious common sense errors. If there are any, deduct 5 points for each error.
4. Understanding of Humor, Creativity, or Magic Effects: Total score of 40 points. If this text explains the counterintuitive phenomenon similar to the main points in the reference answer, give 35 points or above. If it explains the reasons behind the counterintuitive phenomenon but with some differences from the reference answer, give a score between 15 and 35 based on the magnitude of the differences. If this text explains the reasons behind the counterintuitive phenomenon, but the explanation differs significantly from the reference answer, give a score between 0 and 15.
5. Details: Total score of 35 points. In providing the explanation, include specific details about the humorous, creative, or magical aspects of the video. For each additional detail or omission compared to the reference answer, deduct 5 points.
6. If the explanation in this text differs from the reference answer and **includes mostly descriptive details not mentioned in the reference answer**, multiply the total score by 0.5.
7. The lowest possible score is 0, and the highest is 100.

User Input

[It's amusing that a big basin knocked over a child completely, and the child even fell into the basin in the end.]
[The entire video is funny because it depicts the woman using the toilet in the public bathroom, getting stuck, and having to be rescued by the man with the plunger.]

GPT-4 Output

Thank you for providing the texts. Here's how I would evaluate Text 2 according to the provided criteria:

1. **Language Expression (Fluency, Vocabulary):** The text is fluent and uses appropriate vocabulary, providing a coherent scene description. Score: 5/5
2. **Logical Coherence:** The text does not demonstrate logical coherence with the reference answer (Text 1). They appear to describe completely different scenes without any overlapping elements. Score: 0/10
3. **Common Sense Errors:** The text doesn't contain any obvious common-sense errors; it describes a plausible scene. Score: 10/10
4. **Understanding of Humor, Creativity, or Magic Effects:** The explanation in this text differs significantly from the reference answer. There are no points of overlap that would indicate an understanding of the humor, creativity, or magic effects of the reference scene. Score: 0/40
5. **Details:** The text provides specific details, but these details are unrelated to those in the reference answer. The scenes described in each text are entirely distinct. Score: 0/35
6. **Overall adjustment:** The explanation in this text does differ from the reference answer, but instead of adding descriptive details, it depicts a completely different scene. Therefore, no multiplication adjustment is necessary.

Final total score: 5 (Language Expression) + 0 (Logical Coherence) + 10 (Common Sense Errors) + 0 (Understanding of Humor, Creativity, or Magic Effects) + 0 (Details) = 15/100.

Fig. A13: GPT-4 prompt for Explanation Evaluation. We give GPT prompt as system, and then input two explanations, the first one is our annotation, and the second one is the output of the model, and GPT-4 will evaluate the similarity between these two texts according to the prompt's requirements.

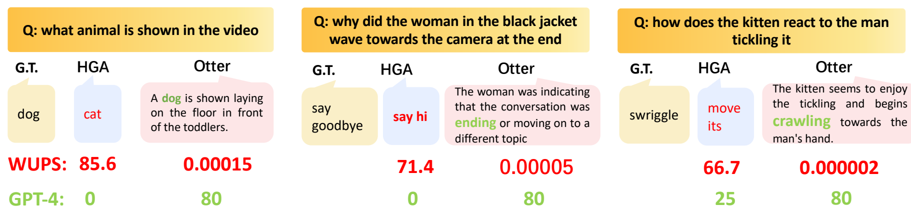


Fig. A14: The responses of VLMs to NExT-OE are scored using the WUPS and GPT-4 based metric. It can be observed that Otter's responses, whether correct or close to the answer, receive scores near zero under the WUPS metric. In contrast, HGA's completely incorrect responses (e.g., answering "cat" when the correct answer is "dog") score highly under WUPS. On the other hand, the scoring by GPT-4 appears to be accurate.

User Input

[It's amusing that a big basin knocked over a child completely, and the child even fell into the basin in the end.]
[The entire video is funny because it depicts the woman using the toilet in the public bathroom, getting stuck, and having to be rescued by the man with the plunger.]

GPT-4 Output 1

Thank you for providing the texts. Here's how I would evaluate Text 2 according to the provided criteria:

1. **Language Expression (Fluency, Vocabulary):** The text is fluent and uses appropriate vocabulary, providing a coherent scene description. Score: 5/5
2. **Logical Coherence:** The text does not demonstrate logical coherence with the reference answer (Text 1). They appear to describe completely different scenes without any overlapping elements. Score: 0/10
3. **Common Sense Errors:** The text doesn't contain any obvious common-sense errors; it describes a plausible scene. Score: 10/10
4. **Understanding of Humor, Creativity, or Magic Effects:** The explanation in this text differs significantly from the reference answer. There are no points of overlap that would indicate an understanding of the humor, creativity, or magic effects of the reference scene. Score: 0/40
5. **Details:** The text provides specific details, but these details are unrelated to those in the reference answer. The scenes described in each text are entirely distinct. Score: 0/35
6. **Overall adjustment:** The explanation in this text does differ from the reference answer, but instead of adding descriptive details, it depicts a completely different scene. Therefore, no multiplication adjustment is necessary.

Final total score: 5 (Language Expression) + 0 (Logical Coherence) + 10 (Common Sense Errors) + 0 (Understanding of Humor, Creativity, or Magic Effects) + 0 (Details) = 15/100.

GPT-4 Output 2

Thank you for providing the texts. Here's how I would evaluate Text 2 according to the provided criteria:

1. **Language Expression (Fluency, Vocabulary):** The sentence is fluent, and the vocabulary is appropriately used. No difficult or overly simplified words were used, and the sentences are clear. However, the phrase "the man with the plunger" could have been expressed in a better way. Still, overall, the language is easily understandable. Score: 4/5
2. **Logical Coherence:** From the given information, it appears that Text2 lacks logical coherence with Text1. The first text talks about a child and a basin, while the second text discusses a woman, a public bathroom, and a man with a plunger. The scenarios are entirely different. Score: 0/10
3. **Common Sense Errors:** There are no obvious common-sense errors. Score: 10/10
4. **Understanding of Humor, Creativity, or Magic Effects:** Given that Text1 is the reference text, Text2 fails to capture the counterintuitive phenomenon described in Text1. The humor and situation in Text2 are entirely different from the reference text. Score: 0/40
5. **Details:** Text2 does not provide any of the details mentioned in Text1. It seems to describe a completely different scenario. Score: 0/35
6. **Overall adjustment:** As the explanation in Text2 differs from Text1 and includes mostly descriptive details not mentioned in the reference answer, the total score should be multiplied by 0.5.

Final total score: 4 (Language Expression) + 0 (Logical Coherence) + 10 (Common Sense Errors) + 0 (Understanding of Humor) + 0 (Details) = 14. Total Score after adjustment = 14 * 0.5 = 7. So, the total score for Text2 based on your criteria is 7/100.

Fig. A15: GPT-4 generates different responses to the same content. User Input is the two sentences input to GPT-4, which is used to judge the similarity. The GPT-4 Output 1 and GPT-4 Output 2 below are the answers given by GPT-4 in two different conversations, and you can see that the two assessments are different.

D.2 On Failure of the Timestamp Localization Task

There is no model that has the ability to answer the Timestamp Localization Task. Fig. A16 shows exactly what is answered for each model.

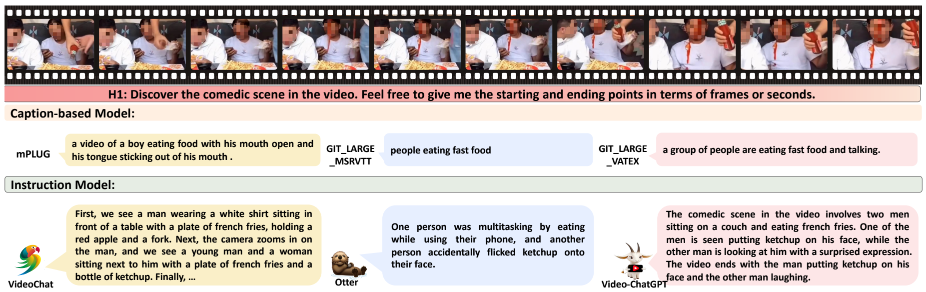


Fig. A16: Model responses on Timestamp Localization Task. As can be seen, the model gives answers to the Timestamp Localization Task still focusing on describing the video content and still does not answer the specific time period when asked to type the number of frames and seconds.

D.3 Implementation Details

mPLUG mPLUG is a multi-modal system employing independent image and text encoders, a cross-modal network, and a text generation decoder, which is trained through prefix language modeling loss to generate captions from connected image and prefix sub-sequence representations.

GIT GIT is a system with an image encoder and a text decoder; it processes multiple video frames independently, adds learnable temporal embeddings before concatenation, uses a contrastively pre-trained model for image encoding, and employs a transformer module for text prediction. We used the 14M version and used two models, GIT_LARGE_VATEX and GIT_LARGE_MSRVTT, which were fine-tuned on the video captioning task for the VATEX and MSRVTT datasets, respectively.

VideoChat VideoChat, specifically the VideoChat-13B version, is an end-to-end system for video comprehension that combines pre-trained models. It utilizes QFormer to generate video embeddings and then employs LLAMA-13B for multimodal understanding and outputs video text descriptions with timestamps. In the experiment, we used VideoChat-13B with the hyperparameters: beam search number = 1, temperature = 1, video segments = 8, and token = 512.

Video-ChatGPT Video-ChatGPT is a vision-language model with a video encoder and LLM. It generates answers using video embeddings and benefits from a data-centric, human-assisted annotation framework for high-quality video instructional data. In the experiment, we used Video-ChatGPT-7B with its hyperparameter: temperature = 0.2, and token = 512.

Otter The Otter model employs the OpenFlamingo training paradigm, utilizing pre-trained encoders for language (LLaMA-7B) and vision (CLIP ViT-L/14). In the fine-tuning process, Otter prioritizes the Perceiver resampler module while keeping the encoders frozen. In the training stage, we finetuned Otter on Dense Caption and FunQA for a total of 3 epochs each. In the experiment, we used two versions of Otter with the same hyperparameters: beam search number = 3, size of no-repeat-ngram = 0.2, and token = 256.

mPLUG-Owl mPLUG-Owl is a novel training paradigm that equips LLMs with multi-modal abilities through modularized learning of foundation LLM, a visual knowledge module, and a visual abstractor module. The training paradigm involves a two-stage method for aligning image and text, which learns visual knowledge with the assistance of LLM, while maintaining and even improving the generation abilities of LLM. In the second stage, languageonly and multi-modal supervised datasets are used to jointly fine-tune a low-rank adaption (LoRA) module on LLM and the abstractor module by freezing the visual knowledge module.

Video-LLaMA Video-LLaMA is built on top of BLIP-2 and MiniGPT-4. It is composed of two core components: (1) Vision-Language (VL) Branch (Visual encoder: ViT-G/14 + BLIP-2 Q-Former) and (2) Audio-Language (AL) Branch (Audio encoder: ImageBind-Huge). In VL Branch, a two-layer video Q-Former and a frame embedding layer (applied to the embeddings of each frame) are introduced to compute video representations. And In AL Branch, a two-layer audio Q-Former and an audio segment embedding layer (applied to the embedding of each audio segment) are introduced to compute audio representations.

D.4 Ablation Experiment about Audio

Audio is undoubtedly an important part of video, but existing models utilize audio in a very limited way, some of them don't use audio as an input, and then get high scores on various lists. For example, Otter, mPLUG, video-ChatGPT and GIT in the paper, the outputs of these models are not affected by the audio, and only the visual information is understood; there are also some models that convert the audio into captions for text encoding and then input them into the model, such as VideoChat in the paper. However, The FunQA dataset is designed to emphasize the understanding of 'fun' through the visual aspect of videos. During the data collection phase, we focused on the visual elements, and in our experimental setup, we retained the audio of the videos when feeding them to the models. To investigate the role of audio, we conducted a set of ablation experiments by muting the audio during both training and testing phases. The results are shown in Table A5. This demonstrates that FunQA is an visual-centric dataset and audio could not ease the challenge of the benchmark.

D.5 More Examples

We conducted experiments to compare machine versus human performance on FunQA. We randomly selected 100 HumorQA pairs from HumorQA and had

Table A5: The ablation experiment about audio. The scores of the results after inputting the muted video, the traditional evaluation scores remain basically unchanged, and the scores of GPT-4 fluctuate very little (this is mainly due to the issues of instability of the evaluation method of GPT-4.

	HumorQA				CreativeQA					MagicQA		
	H1	H2	H3	H4	C1	C2	C3	C4	C5	M1	M2	M3
VideoChat [17]	-	0.5 / 13.7 / 0.0 44.0 / 37.9	0.5 / 13.5 / 0.0 45.4 / 31.9	0.8 / 5.1 / 0.5 20.2 / 61.7	-	0.3 / 7.5 / 0.0 21.7 / 10.9	0.3 / 7.7 / 0.0 22.8 / 27.7	0.2 / 1.2 / 0.2 7.3 / 51.1	67.5	-	0.6 / 15.5 / 0.0 47.4 / 14.2	0.3 / 9.2 / 0.0 43.1 / 24.6
VideoChat (MUTED) [17]	-	0.5 / 13.5 / 0.0 40.0 / 34.5	0.5 / 12.5 / 0.0 43.7 / 31.9	0.8 / 5.1 / 0.5 22.4 / 57.8	-	0.3 / 7.4 / 0.0 24.0 / 10.0	0.4 / 7.7 / 0.0 20.2 / 30.5	0.2 / 1.2 / 0.2 6.8 / 45.3	67.5	-	0.6 / 15.0 / 0.0 48.0 / 12.2	0.3 / 9.2 / 0.0 44.0 / 23.4

Table A6: Human Performance on FunQA. 100 QA pairs (H2, H3, H4) from HumorQA in FunQA were randomly selected and answered by five different people. We use the following metrics: **BLEU-4 / ROUGE-L / CIDEr / BLEURT / GPT-4** for evaluation.

Task	H2	H3	H4
Human 1	6.0 / 28.8 / 6.5 / 46.6 / 82.4	2.2 / 20.8 / 1.1 / 45.0 / 85.2	0.8 / 9.2 / 1.1 / 20.5 / 76.6
Human 2	13.8 / 42.5 / 11.1 / 55.4 / 90.5	13.2 / 34.1 / 10.0 / 54.4 / 87.5	1.1 / 12.8 / 2.4 / 33.5 / 80.6
Human 3	0.7 / 16.5 / 0.5 / 36.2 / 77.5	0.7 / 14.7 / 0.2 / 44.1 / 69.0	0.2 / 3.1 / 0.3 / 15.4 / 76.8
Human 4	26.4 / 50.5 / 27.7 / 75.8 / 80.3	10.3 / 36.8 / 11.5 / 74.3 / 90.0	2.8 / 21.7 / 6.0 / 43.4 / 73.4
Human 5	3.5 / 16.7 / 3.7 / 48.0 / 86.6	2.4 / 16.4 / 1.8 / 47.9 / 90.3	0.8 / 10.5 / 0.8 / 23.9 / 88.2
Model (SOTA)	1.5 / 18.1 / 1.0 / 44.0 / 37.9	1.3 / 15.4 / 0.5 / 25.9 / 61.7	**1.1** / 8.8 / 0.7 / 25.9 / 61.7
Human (Avg)	**10.1** / **31.0** / **9.9** / **60.0** / **83.6**	**5.8** / **24.6** / **5.0** / **53.1** / **84.4**	**1.1** / **11.46** / **2.1** / **27.3** / **79.1**

five individuals who had not watched any FunQA videos provide answers to the questions (H2, H3, H4) for each video. Subsequently, we calculated the average score for each individual under the FunQA benchmark and the overall average score of human responses. The FunQA benchmark score is the selected SOTA score between models.The results are in Table A6.

D.6 More Examples

Fig. A17 shows the responses of different models on CreativeQA and MagicQA.

E More Discussions

E.1 (Potential) More Essential Factors

Accurate understanding of the videos Through our analysis of failure cases, we've observed that many models struggle with accurately describing videos. While they might be adept at detecting objects within the videos, they often falter in comprehending the contextual relationship between sequential events. Such misinterpretations indicate that there's a need for further exploration in this domain. The videos we've used can indeed serve as an invaluable dataset for probing video descriptions in depth.

Logic Reasoning The primary nature of our videos encompasses content that is counterintuitive and contradicts common sense. For models to understand

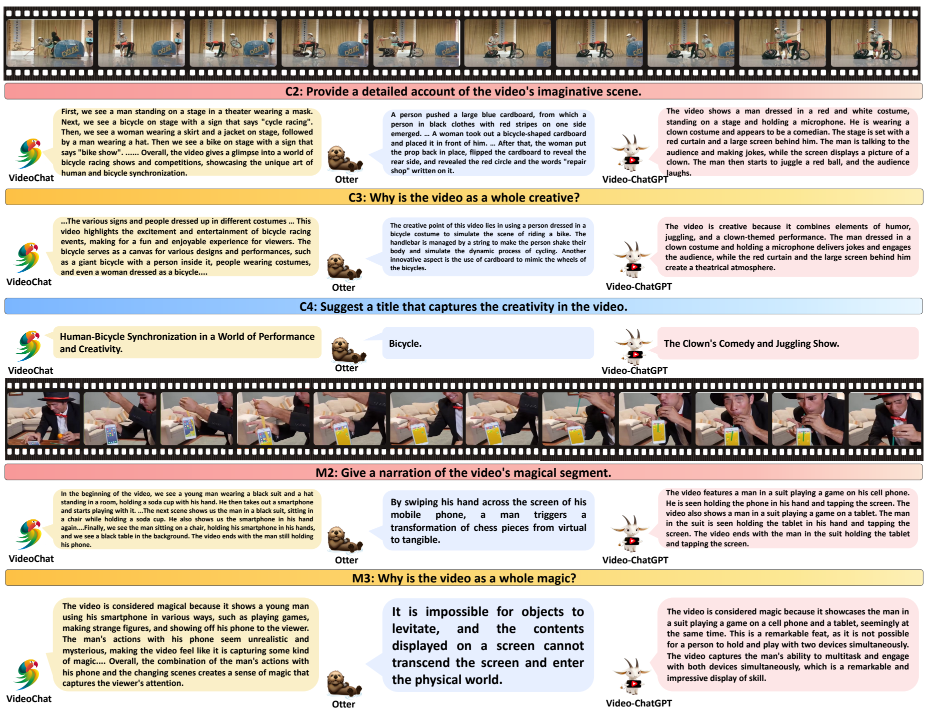


Fig. A17: Model responses on CreativeQA and MagicQA. For the description of the Creative video example, only VideoChat gives the key point of the bicycle, but its description also has many errors and omissions, and the remaining two models do not identify the bicycle. In the explanation task, the responses of all three models fail to clearly explain the creativity of this imitation performance. For the Magic video example, all three models perform very poorly in description and explanation, basically only answering the phone and the straw, but lacking the description and explanation of the magic effect.

these, it's imperative they grasp the concept of "common sense." They must deduce what would typically transpire under normal circumstances and then use that perspective to humorously interpret the video. This necessitates the model to possess strong reasoning capabilities, especially when it comes to common sense reasoning.

Extra Knowledge - Sense of Humor To decipher the humor in a video, it's plausible that understanding the fundamental principles of humor is crucial. This type of knowledge, along with many other tidbits of common sense and additional information, might enhance the model's performance. Determining how to integrate valuable knowledge and discerning what counts as "valuable" are topics that warrant further exploration.

E.2 Potential Solutions

Model Size Increasing the number of parameters is a natural method to enhance the model's performance. However, this approach comes with its own set of engineering challenges, requiring improvements in model optimization and deployment. We're also curious about the relationship between the number of parameters and the performance on FunQA. This is an intriguing research point in itself, and our dataset can serve as an excellent test bed to further this exploration.

Data Quality We believe the emphasis for this task should be on data collection. Current trends with large and dynamic models have shown that having vast amounts of low-quality data isn't as effective as a smaller quantity of high-quality data. Thus, we hope the community can discover the type of data that genuinely assists in understanding counterintuitive videos. This is a crucial research direction.

Training Strategy Studying training strategies is also essential. For instance, determining which type of data to start learning from, and understanding the significance of curriculum learning, among others.

Model Collaboration Ultimately, we might not need to solely focus on a single model to solve this problem. Perhaps multiple models collaboratively working on examples in an elegant manner could be a method to enhance performance. However, this approach might necessitate paying more attention to the overall efficiency of model implementation

E.3 The Emphasis of Temporal Dynamics in FunQA

HumorQA Example: Humor_Example.mp4

H2: An individual slipped on the staircase filled with ice and tumbled down to the very bottom, followed by a second person who also fell after witnessing the first person's fall.

H3: The first person falling down the stairs step by step was already very funny, and the second person repeating the same mistake and falling down made it even more hilarious.

Explanation: As answered in H2, the main element of humor in this video is two people slipping down an icy staircase in a sequential order, and if the model does not make sense of the temporal information, it will not be able to give the sequential logic of the two people slipping down one after the other. It is also clear from the answer to H3 that the sequence of slipping backwards and forwards is one of the sources of humor, as the person at the back does not learn from the lesson of the former and slips down in a similar way, and this repetition of the wrongdoing brings humor to the situation.

CreativeQA Example: Creative_Exapmle.mp4

C2: At the center of the stage is a blue rectangular box. Following the little girl's watering of the box, four individuals lying inside gradually lift their hands and legs, each at different paces and heights, until they stop and reveal green painted leaf-shaped objects on their limbs.

C3: The main creative element of this video is when the four individuals in a box, with varying heights and movement speeds, gradually raise their hands and feet to mimic the growing process of the carrot seeds planted by the little girl. The green leaf-shaped objects, tied and opened in advance with their hands and feet, are used to simulate the true sprouting of carrot sprouts, resulting in a lively and imaginative scene.

Explanation: The performance in this video mimics growing bean sprouts, the process of growing beans into sprouts by constantly watering the soil. Trying to describe the content of this video idea and explain where the idea came from requires an understanding of the temporal information. Analyzing the visual information together, the model has to understand the sequence of watering, bean sprouts growing from the soil, and bean sprouts growing taller and greener in order to answer the question accurately. This difference in understanding can be demonstrated on the C3 task, where without analyzing the temporal information, the optimal answer would be bean sprouts growing from the soil, and cause and effect logical relationships such as watering to grow bean sprouts would be ignored.

References

1. Boyd, B.: Laughter and literature: A play theory of humor. Philosophy and literature **28**(1), 1–22 (2004) 1
2. Castro, S., Wang, R., Huang, P., Stewart, I., Ignat, O., Liu, N., Stroud, J.C., Mihalcea, R.: Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework (2022) 2
3. Corporation, N.T.N.: Kasou taishou. `https://www.ntv.co.jp/kasoh/index.html`, [Accessed 23-Apr-2023] 1
4. Gao, D., Wang, R., Bai, Z., Chen, X.: Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. pp. 1675–1685 (October 2021) 2
5. Garcia, N., Otani, M., Chu, C., Nakashima, Y.: Knowit vqa: Answering knowledge-based questions about videos. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (2020) 2
6. Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Agqa: A benchmark for compositional spatio-temporal reasoning (2021) 2
7. Jang, Y., Song, Y., Kim, C.D., Yu, Y., Kim, Y., Kim, G.: Video Question Answering with Spatio-Temporal Reasoning. IJCV (2019) 2
8. Kant, I.: Critique of judgment. Hackett Publishing (1987) 1
9. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017) 4
10. Koestler, A.: The act of creation. In: Brain Function, Volume IV: Brain Function and Learning, pp. 327–346. University of California Press (2020) 1
11. Lamont, P., Wiseman, R.: Magic in theory: An introduction to the theoretical and psychological elements of conjuring. Univ of Hertfordshire Press (2005) 2
12. Latta, R.L.: The basic humor process: A cognitive-shift theory and the case against incongruity. De Gruyter Mouton (1999) 1
13. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: EMNLP (2018) 2
14. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvqa+: Spatio-temporal grounding for video question answering. In: Tech Report, arXiv (2019) 2
15. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 12
16. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., Si, L.: mplug: Effective and efficient vision-language learning by cross-modal skip-connections (2022) 12
17. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023) 12, 19
18. Morreall, J.: Philosophy of Humor. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Summer 2023 edn. (2023) 1
19. Muhammad Maaz, Hanoona Rasheed, S.K., Khan, F.: Video-chatgpt. `https://github.com/mbzuai-oryx/Video-ChatGPT` (2023) 12
20. Mun, J., Hongsuck Seo, P., Jung, I., Han, B.: Marioqa: Answering questions by watching gameplay videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2867–2875 (2017) 2

21. Nelms, H.: Magic and showmanship: A handbook for conjurers. Courier Corporation (2012) 2
22. Runco, M.A., Jaeger, G.J.: The standard definition of creativity. CREATIVITY RESEARCH JOURNAL **24**(1), 92–96 (2012) 1
23. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016) 2
24. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022) 12
25. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: STAR: A benchmark for situated reasoning in real-world videos. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021), https://openreview.net/forum?id=EfgNF5-ZAjM 2
26. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa:next phase of question-answering to explaining temporal actions (2021) 2
27. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017) 2
28. Xu, L., Huang, H., Liu, J.: Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9878–9888 (2021) 2
29. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 4
30. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: Avqa: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3480–3491 (2022) 2
31. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Jiang, C., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., Huang, F.: mplug-owl: Modularization empowers large language models with multimodality (2023) 12
32. Ye, Y., Zhao, Z., Li, Y., Chen, L., Xiao, J., Zhuang, Y.: Video question answering via attribute-augmented attention network learning. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17, ACM (Aug 2017). https://doi.org/10.1145/3077136.3080655, http://dx.doi.org/10.1145/3077136.3080655 2
33. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning (2020) 2
34. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI. pp. 9127–9134 (2019) 2
35. Zadeh, A., Chan, M., Liang, P.P., Tong, E., Morency, L.P.: Social-iq: A question answering benchmark for artificial social intelligence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8807–8817 (2019) 2

36. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023), https://arxiv.org/abs/2306.02858 12