FunQA: Towards Surprising Video Comprehension

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

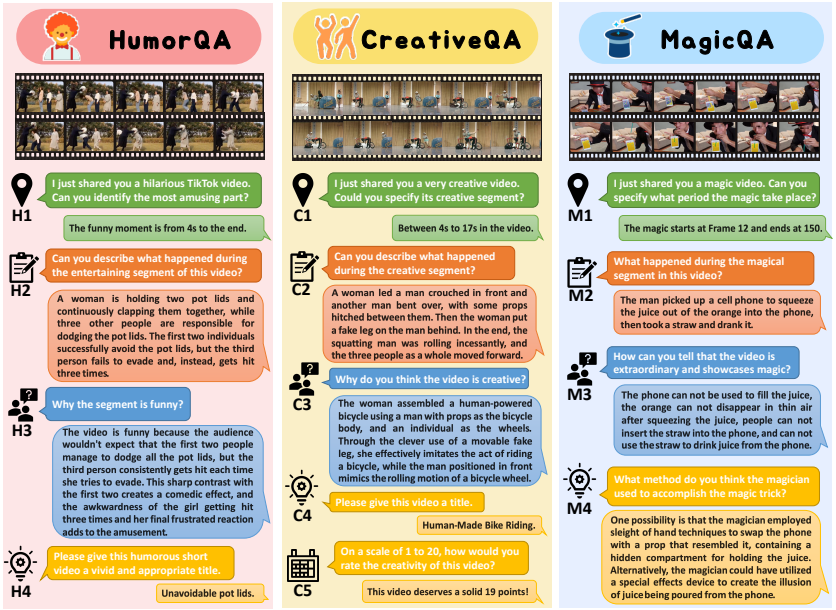

Fig. 1: Overview of FunQA. FunQA comprises three subsets of surprising videos: *1) HumorQA, 2) CreativeQA,* and *3) MagicQA*. Each subset is associated with three common tasks: *1) counter-intuitive timestamp localization*, *2) detailed video description*, and *3) reasoning around counter-intuitiveness* (see **H1-3**, **C1-3**, and **M1-3**). Furthermore, we offer higher-level tasks tailored for each video type, such as *attributing a fitting and vivid title* for HumorQA and CreativeQA (see **H4**, **C4**), etc.

Abstract. Surprising videos, *e.g.*, funny clips, creative performances, or visual illusions, attract significant attention. Enjoyment of these videos is not simply a response to visual stimuli; rather, it hinges on the human capacity to understand (and appreciate) commonsense violations depicted in these videos. We introduce **FunQA**, a challenging video question answering (QA) dataset specifically designed to evaluate and enhance

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

the depth of video reasoning based on counter-intuitive and fun videos. Unlike most video QA benchmarks which focus on less surprising contexts, e.g., cooking or instructional videos, FunQA covers three previously unexplored types of surprising videos: **1) HumorQA**, **2) CreativeQA**, and **3) MagicQA**. For each subset, we establish rigorous QA tasks designed to assess the model's capability in counter-intuitive timestamp localization, detailed video description, and reasoning around counter-intuitiveness. We also pose higher-level tasks, such as attributing a fitting and vivid title to the video, and scoring the video creativity. In total, the FunQA benchmark consists of 312K free-text QA pairs derived from 4.3K video clips, spanning a total of 24 video hours. Moreover, we propose **FunMentor**, an agent designed for Vision-Language Models (VLMs) that uses multi-turn dialogues to enhance models' understanding of counter-intuitiveness. Extensive experiments with existing VLMs demonstrate the effectiveness of FunMentor and reveal significant performance gaps for the FunQA videos across spatial-temporal reasoning, visual-centered reasoning, and free-text generation.

1 Introduction

The charm of surprising videos, being funny, creative, and filled with visual illusions, offers enjoyment and attracts engagement from viewers. This type of media elicits *positive surprise*[1] [50], a captivating emotion that stems not merely from perceiving surface-level visual stimuli, but rather, the innate ability of humans to understand and find delight in unexpected and counter-intuitive moments [44]. However, despite significant advancements in today's computer vision models, the question remains: can video models "understand" the humor/creativity in surprising videos? Consider the humorous video depicted in Fig. 1 (left) as an example. We observe a woman in black holding two pot lids and clapping them together. The remaining three individuals are responsible for avoiding the pot lids. The first two people successfully dodge, but the third girl, in a panic, fails to avoid any hits and gets struck three times. The embarrassed demeanor of the third girl along with her final frustrated reaction, elicits laughter[2]. While humans effortlessly recognize this as an unusual (and potentially entertaining) event, the reasoning required to holistically understand the scene is complex: a model needs to recognize that this is not a video depicting harm but rather girls *engaging in playful pranks together*, and discern that the comedic element arises from *the stark contrast between the third girl being hit by the pot lids every time and the first two girls skillfully avoiding them.*

While there have been some efforts to enhance computer vision models' performance in Video Question Answering (VideoQA), these works have primarily focused on the common, less surprising videos found in existing VideoQA datasets.

[1] c.f., *negative* surprise, e.g., a surprising medical bill.

[2] The hostility/superiority theory of humor posits that humor can arise from claiming superiority over someone or something [2,17]; but alternate (more optimistic) theories of humor exist, [1] offers a survey.

Table 1: Comparison between FunQA and other existing benchmarks. 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 [20]	Social Media	🤖	3s	72K	165K	✓	✓	✗	✓	✗	2.1
MSRVTT-QA [64]	Social Media	🤖	15s	10K	244K	✓	✗	✗	✓	✗	1.0
ActivityNet-QA [72]	Social Media	👨‍🎓	180s	6K	58K	✓	✗	✗	✓	✗	1.9
NExT-QA [63]	Daily life	👨‍🎓	44s	5K	52K	✓	✓	✓	✓	✓	2.6
Social-IQ [73]	Daily life	👨‍🎓	99s	1K	8K	✓	✗	✓	✗	✓	N/A
MovieQA [58]	TV shows	🤖	203s	7K	6K	✗	✗	✓	✓	✓	N/A
TVQA+ [33]	TV shows	👨‍🎓	8s	4K	30K	✗	✗	✓	✓	✓	N/A
SUTD-TrafficQA [65]	Traffic	👨‍🎓	5s	10K	623K	✓	✗	✗	✓	✓	N/A
MarioQA [48]	Games	👨‍🎓	5s	188K	188K	✓	✗	✓	✓	✗	2.0
CLEVRER [71]	Synthetic Videos	👨‍🎓	5s	20K	305K	✓	✗	✓	✓	✓	N/A
FunQA (Ours)	**Surprising Videos**	👨‍🎓	19s	4K	**312K**	✓	✓	✓	✓	✓	**34.2**

Examples of commonly employed VideoQA datasets include YouCook2 [76] which contains video clips from 2K cooking videos, Howto100M [45] which consists of only instructional videos. While there exist video datasets that explore the humor in TV shows [5, 18] and include tasks such as predicting laughter tracks [53], these tasks often heavily rely on audio and narrative cues, with visual clues might playing a lesser role. Beyond datasets centered on factual queries, it is worth noting that NExT-QA targets the explanation of video content, which is widely employed for evaluating reasoning abilities. However, it was found in the experiment (see Section 5.4) that VLMs such as GPT-4V(ision) already achieved an accuracy of 80% on NExTQA. This demonstrates that with the development of VLMs, the demand for datasets with deeper reasoning capabilities and presenting greater challenges is increasing.

To revitalize the visual reasoning field and further improve model capabilities to identify and understand visual commonsense violations in videos, we introduce **FunQA**[1], an extensive and carefully curated VideoQA dataset comprising 4.3K surprising videos and 312K manually annotated **free-text** QA pairs. Unlike some VideoQA datasets that feature open-ended questions but short answers (e.g., an average of 2.6 words per answer in NExT-QA [63]), FunQA's responses average 34.2 words in length. This significantly increases the demand for advanced video comprehension capabilities in the model. Therefore, here we use **free-text** QA to distinguish from open-ended QA. Our dataset consists of three subsets: **1) HumorQA**, **2) CreativeQA**, and **3) MagicQA**. Each subset covers different sources and contents, but the commonality lies in their surprising nature, e.g., the unexpected contrasts in humorous videos, the intriguing disguises in creative videos, and the seemingly impossible performances in magic videos.

[1] FunQA has been integrated in LMMs-Eval [34], where can easily obtain FunQA dataset and evaluate multiple VLMs.

Our experiments suggest that these surprising videos require different types of reasoning than common videos, as existing VideoQA methods perform poorly on the corpus. With FunQA, we hope to provide a benchmark that covers the popular, important, and sophisticated genre of counter-intuitive/surprising videos.
In FunQA, we formulate three rigorous tasks to measure models' understanding of surprise: **1) Counter-intuitive timestamp localization**: a model must identify the specific time period within a video when an unexpected event takes place. **2) Detailed video description**: a model must generate coherent and objective descriptions of the video content, evaluating models' fundamental video understanding capabilities. **3) Counter-intuitiveness reasoning**: a model must generate concrete explanations of why the video is surprising. These tasks progressively assess the model's ability to perceive, articulate, and reason about the counter-intuitive elements present in surprising videos. We also propose additional tasks that pose higher-level challenges, such as assigning an appropriate and vivid title to the video.
In continuation of our efforts to enhance models' comprehension of surprising content, we introduce **FunMentor**, a specialized agent designed to boost counter-intuitive reasoning in VLMs. Operating like a seasoned coach in a variety show, FunMentor engages in detailed, multi-turn dialogues, honing the models' responses to accurately grasp the essence of both amusing and astonishing content. FunMentor actively steers VLMs with precise prompts, fostering fluent, logical, and persuasive responses. Experiments have demonstrated its effectiveness in augmenting VLMs' ability to comprehend. To summarize our contributions:
1) New VideoQA Dataset: We build a large-scale dataset **FunQA**, which complements the existing VideoQA dataset with intriguing videos.
2) Novel and Challenging Tasks: We design a number of novel tasks that allow the model to explore previously untouched problems, such as timestamp localization, and reasoning around counter-intuitiveness. These tasks push video reasoning beyond superficial descriptions, demanding deeper understanding.
3) Novel Method FunMentor: We propose a novel agent refines the model's understanding of counter-intuitiveness through multi-turn dialogues with VLMs.
4) Comprehensive Evaluation: We have done an comprehensive evaluation of cutting-edge baselines, giving the field an insight and future research direction.

2 Related Work

Video Question Answering Benchmarks While the visual question answering (VQA) task focuses on enhancing models' ability in image comprehension [15, 28, 77], video question answering (VideoQA) shifts the attention towards video comprehension. VideoQA is generally more challenging than VQA as it requires a comprehensive understanding of visual content, utilization of temporal and spatial information, and exploration of relationships between recognized objects and activities [71]. To address the VideoQA task, the research community has introduced various benchmarks. As depicted in Table 1 (the complete table are shown in Appendix A), most commonly used VideoQA datasets are sourced

from human-centric videos like movies [58], TV shows [14, 32, 33], and social media [6, 16, 20, 62, 63, 68, 73], and there are also object-centric datasets of game videos [48], synthetic videos [71] and egocentric videos [12]. MovieQA [58] and TVQA [32] are commonly employed by VideoQA methods, which put forward tasks related to temporal and causal reasoning. However, they rely heavily on dialogue comprehension and textual plot summaries, which severely limits the challenge of visual reasoning. TGIF-QA [20] uses animated GIFs to challenge spatial-temporal reasoning, but as most GIFs are short videos of 3 seconds, and its tasks mainly focus on action description, TGIF-QA lacks complex reasoning evaluation ability. When most datasets use multiple choice questions as QA tasks, some methods, such as NExT-QA [63], try to join open-ended questions. NExT-QA mainly focuses on daily life videos, but the open-ended answers are mostly simple sentences containing only a few words. To sum up, most existing methods focus on ordinary videos, lack of understanding of intriguing or unexpected videos, and advanced reasoning tasks such as generating complete explanatory texts of videos remain to be explored.

Video Question Answering Methods Earlier studies have explored various models, including LSTMs and graph-based neural networks, to capture cross-modal information [38, 75]. With the advent of Transformers, video understanding models, like ClipBERT [31] and CoMVT [57] emerged, focusing on understanding specific frames within a video. Subsequent models like Violet [11], extended their ability to encompass temporal and spatial information. However, these methods have primarily been applied to short videos. For long videos, MIST [13] stands out by achieving state-of-the-art (SOTA) performance and excelling in terms of computation efficiency and interpretability. Furthermore, recent VLMs [35, 37, 47] have showcased remarkable video understanding capabilities.

Counter-Intuitive Benchmarks While many current computer vision benchmarks primarily focus on understanding commonsense content, there is a growing interest in addressing the realm of counter-intuitiveness. Several emerging benchmarks and models cater to this domain, such as Whoops [3], which emphasizes weird, unusual, and uncanny images, OOPS [9], which centers on recognizing and predicting unintentional events, and MemeGraphs [26], which revolves around memes featuring humor and sarcasm. Furthermore, some works even challenges models to comprehend complex multimodal humor in comics [19]. In the realm of large language models, exemplified by GPT-4 [51], there is a particular focus on showcasing their ability to provide explanations for funny pictures. However, regarding videos, existing datasets exploring humor in TV shows or comedy tend to heavily rely on audio and narrative cues [5, 18, 53], with visual clues playing a comparatively lesser role.

3 The FunQA Dataset

In this section, we provide a detailed explanation of the design principles that guided the creation of the FunQA dataset and its subsets. Additionally, we introduce our novel VideoQA tasks tailored for FunQA, and FunQA data statistics

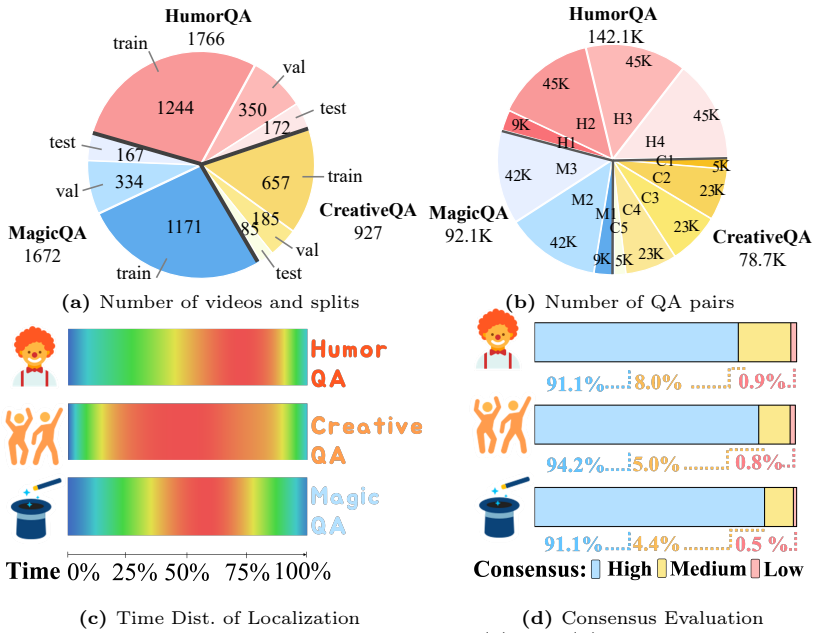

Fig. 2: Statistics of FunQA Dataset. Figure (a) and (b) showcase vital statistics, including the number of videos for different videos, splits, and QA pairs count for three subsets. Figure (c) highlights the high-frequency time span of the answer for localization questions in red. Figure (d) presents the percentage of consensus between annotators for the same QA pair. The consensus is categorized into three levels **High** for consistent understanding, **Medium** for partial agreement but with mutual acknowledgment, and **Low** for complete disagreement.

in Figure 2. Toward the end we present our Construction Pipeline and Quality Control, highlighting our efforts in maintaining data quality and objectivity.

3.1 Task Definition

To comprehensively evaluate the model's ability to understand surprising videos, we designed the following 4 types of tasks for each subset:

Counter-intuitive Timestamp Localization Task The localization task is the base task to assess the model's comprehension abilities. It involves localizing counter-intuitive segments within the video, answers expressed in either seconds or frames. This task serves as the basis for the subsequent two main tasks in the three subsets, where the focus shifts to locating moments of humor, creativity, and magical effects, respectively. Successfully completing this task demands the model's understanding of the video's overall content, incorporating both temporal and spatial information.

Detailed Description Task The description task aims to evaluate the model's information extraction capabilities, serving as a fundamental aspect of video understanding. This task requires providing a free-text answer that describes the selected moment. Furthermore, this task allows for analysis of how the model

extracts information and generates answers for subsequent tasks. By examining the model's performance in this task, we gain insights into its ability to extract relevant information and generate meaningful responses.

Counter-intuitiveness Reasoning Task The reasoning task is designed to test the model's ability to reason about the video, and in the three subsets, this question is Why Humorous, Why creative, and Why counter-intuitive and the answer is a free-text explanation. This task is very difficult and involves the model's deep reasoning ability; it requires the model to give a complete explanation using information from the entire video and its own common sense.

Higher Level Tasks In addition to the three main tasks, we design higher-level tasks to enhance the model's inference abilities on counter-intuitive videos. *Title Task* in HumorQA and CreativeQA requires generating a concise title summarizing the video's content. *Creative Scoring Task* in CreativeQA involves rating the creativity of videos between 1 and 20 (provided officially). *Magic Method Task* in MagicQA requires the model to explain clearly the rationale behind the magic, and its purpose is to test the model's ability to reason more deeply. To ensure the accuracy of the answers, this task is only partially annotated and appears only in the test set, details of which can be found in Appendix A.1.

3.2 Dataset Statistics

FunQA contains **4,365** counter-intuitive video clips and **311,950** QA pairs, the total length of these videos is **23.9h** and the average length of each clips is **19** seconds. FunQA consists three fine-grained subsets, each one containing well-designed tasks. The specific numbers of videos and splits can be seen in Fig. 2 (a). The specific number of QA pairs for each task can be seen in Fig. 2 (b). For our localization task, the heat map for the three different types of videos can be seen in Fig. 2 (c), which shows the high-frequency time span of the answer. For the description and reasoning tasks, the average length of the words in their free-text answers reached **34.24**, which is much longer than existing VideoQA datasets (e.g., 2.6 in NExT-QA [63]). FunQA has a well-established annotation process and high annotation quality, the result of our annotation consensus evaluation are illustrated in Fig. 2 (d). Impressively, over 90% of the annotations demonstrate a high level of consensus, while only 1% exhibit low consensus. This clearly underscores the objectivity and reliability of the FunQA dataset.

3.3 Dataset Construction Pipeline

FunQA dataset construction pipeline was in three stages: Pre-processing, Manual Annotation, and Post-Processing. The whole process took about 900 hours with over 50 highly educated undergraduates as part-time annotators. See Appendix A.1 for more details on the dataset construction.

Pre-Processing Initially, we crawled videos from YouTube. Then we performed a two-stage manual cleaning and cutting process on the collection to ensure counter-intuitive features and video quality and to exclude non-ethical and sensitive content, resulting in video clips.

Manual Annotation We annotated the videos according to the characteristics of different task designs in Chinese. We screen and train the annotators to ensure the accuracy and high quality of the annotation, and finally produce the original annotated files. After the first round, we conducted a secondary round of 10% of the tasks and performed Consensus Evaluation to ensure the objectivity.

Post-Processing Based on our carefully designed tasks and high-quality annotations, we expanded our dataset using GPT-3.5. Firstly, we automatically translated the Chinese annotations into English. Subsequently, we generated more QA pairs that were faithful to the original ideas but presented differently. This not only made FunQA multilingual but also expanded its QA pair count to 312K. Additionally, we created diverse sub-dataset, FunQA-MC (multi-choice QA) and FunQA-DIA (dialogue QA). In addition, to focus on exploring the ability to handle counter-intuitive reasoning, we released FunQA-MC-R (a multi-choice version specifically containing counter-intuitive reasoning questions). More details are given in Appendix A.2 and Appendix B.

Table 2: Consensus Evaluation Experiment. This table shows the results from a random 10% sample of QA pairs, cross-validated by annotators to assess agreement with existing annotations. 'Low,' 'Medium,' and 'High' indicate the strength of the consensus.

Consensus	HumorQA	CreativeQA	MagicQA	# Total
Low	1	1	2	4 **(1%)**
Medium	9	6	17	32 **(8%)**
High	199	111	194	504 **(91%)**

Table 3: Can FunQA be Solved Solely Based on Images? The left two columns show the average number of questions answerable or not by humans using only 8 static, uniformly selected frames.

Dataset	# Can	# Cannot	Cannot Rate
HumorQA	7.8	32.2	80%
CreativeQA	6.1	33.9	85%
MagicQA	2.6	37.4	94%
FunQA	16.5	103.5	**86%**

3.4 Quality Control

Minimal Errors and High Objectivity in FunQA We assure that every annotation included in the final release of FunQA has been subjected to rigorous **multi-person**, **multi-round** review processes. Furthermore, We did manual consensus evaluation on released FunQA dataset, randomly sampling 10% of the data from all three sub-datasets (HumorQA, CreativeQA, and MagicQA). As shown in Table 2, we get the 91% high consensus.

FunQA Emphasis on Temporal Dynamics FunQA requires a strong emphases on temporal dynamics rather than solely on few frames of images. To prove that, we did the quantitative human experiments - We randomly selected 40 videos from each of the three sub-datasets, totaling 120 videos. For each video, we sampled 8 frames evenly. We enlisted 10 individuals who had not seen any FunQA videos before. We had them view the sequence of 8 consecutive frames and then watch the original video along with its annotations. They were asked to determine whether they can understand and answer the counter-intuitive understanding of the original video solely based on the images. Nearly **86%** people thought that FunQA cannot be solved only by images, as shown in Table 3.

4 FunMentor

This section presents the details of **FunMentor** for counter-intuitiveness understanding. FunMentor is an agent that refines a VLM's answer through multi-turn dialogues, ensuring it generates answers that best explain the given surprising content. The refining process comprises three components:

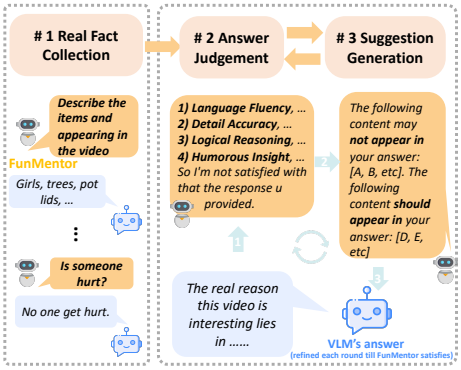

Fig. 3: FunMentor's Refining Process. FunMentor asks multi-round questions to help VLMs to generate persuading answers.

Real Fact Collection Initially, FunMentor poses a series of inquiries to the VLM model it aims to assist, focusing on fact-aware questions to accurately comprehend the objective content of the video (e.g., FunMentor might ask: "Please describe the items and characters appearing in each frame of the video"). This step is crucial because, without this preliminary context, FunMentor would have no knowledge of the video content and might be gullible to VLM's humorous explanations of the video. Thus, providing it with some factual clues is very important.

Answer Judgement FunMentor assesses VLMs' responses based on several key aspects: **1) Language Fluency**, ensuring responses are grammatically fluent; **2) Detail Accuracy**, verifying factual correctness in relation to the video's content; **3) Logical Reasoning**, checking for coherent logic and smooth transitions; and **4) Humorous Insight**, expecting responses to provide humor beyond a mere description of the video. If not passed, FunMentor will try to generate some feedback, which is explained below.

Suggestion Generation Upon receiving unsatisfactory responses, FunMentor formulates constructive suggestions and tailored prompts, which are then sent back to the VLMs for revision, anticipating their improved answers. This process involves the agent analyzing the model's initial response, blending pre-defined prompts with the context of the original question, to generate specific feedback instructions. For example, FunMentor examines the issues in the VLM's response and advises on what should and shouldn't be included in the revised answer, guiding the VLM towards a more accurate and relevant reply.

Fig. 3 shows the pipeline of FunMentor. More details as shown in Appendix C.

5 Experiments

In this section, we begin by introducing various caption-based and instruction-based models for evaluation. We then delve into diverse metrics tailored for FunQA tasks, with extensive details provided in Appendix D. Our comprehensive experiments and deep analysis of the results are then presented. Additionally, we conduct a comparative study contrasting FunQA with the previous VQA dataset,

NExT-QA, which similarly features multiple-choice and open-ended questions. This comparison underscores the unique attributes and importance of FunQA.

5.1 Model Zoo

We categorize the models capable of addressing the majority of VQA tasks into two classes: Caption-based and Instruction-based. Caption-based models generate suitable captions for videos by taking different prompts as input. For this category, we evaluate mPLUG [36] and GIT [61]. Meanwhile, Instruction-based models are capable of answering a wide array of questions. For this category, we evaluate models VideoChat [37], Video-ChatGPT [47], mPLUG-Owl [69], Video-LLaMA [74], Otter [35], Video-LLaVA [39] and LLaVA-NeXT [42]. For Otter, we evaluate two versions: one is fine-tuned on the Dense Caption [27], and another is fine-tuned on the FunQA training set. We evaluate the proposed FunMentor based on Video-ChatGPT and Otter.

5.2 Evaluation Metrics

Timestamp Localization (H1, C1, M1) We employ the intersection of unions (IOU) based on time span.
Description & Reasoning (H2-4, C2-4, M2-3) For all the **free-text** tasks, we employ two approaches for evaluation. Firstly, we utilize traditional NLG (Natural Language Generation) metrics. We use BLEU-4 [52], ROUGE-L [40], CIDEr [59], and BLEURT [54] as our metrics. The first two rely on N-gram overlap, which is only sensitive to lexical variations and cannot identify changes in sentence semantics or grammar. The latter two are reference-based evaluation metrics. Secondly, several works [7, 10, 22, 60] have shown promising results in utilizing GPT as a metric for NLG. Therefore, we introduce GPT-4 to assist in evaluating free-text similarity. We carefully design the prompts to make it possible to give objective ratings as much as possible like a human being. More details of GPT-4 prompts and evaluation criteria are provided in Appendix D.1.
Creative Scoring (C5) $CreativeScoreMetric = 100 \times \left(1 - \frac{|Predict - GT|}{20}\right)$.

5.3 Results and Observations

In Table 4, we show the results of model zoo and our proposed method. For clarity, we provided a list of dos and don'ts for comparing values in the table:
a. Values from different tasks with the same video type (e.g., H2 and H3) are not comparable. We observe that the model output in reasoning tasks may contain several words that match the ground truth (GT) while the meaning might be incorrect, resulting in inflated results. Therefore, our comparison is only based on qualitative analysis.
b. Same metric with different models (see vertically) is comparable.
c. Same task for different video types (e.g., H2 and C2) is comparable. As an example, Fig. 4 illustrates the responses of VLM before and after applying

Table 4: Main Results on FunQA Benchmark. The FunQA benchmark consists of four task categories. **H1, C1, M1** represent the counter-intuitive timestamp localization task, where **IOU** is used as the metric. **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 utilize the **BLEURT** (B.) and **GPT-4** metrics for evaluation. The scores for the traditional metrics (BLEU-4 and CIDEr) are all close to zero. Here, we present only the BLEURT and GPT-4 scores, with the full results available in the Appendix D.3). **C5** represents scoring the video creativity, and the metric is the **Accuracy** between the predicted score and the official score. Here we clarify the abbreviation in the table: **F** denotes Frame-rate; **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	H1	H2-Des.	H3-Rea.	H4-Title	C1	C2-Des.	C3-Rea.	C4-Title	C5-Score	M1	M2-Des.	M3-Rea.
Metrics	IOU	B. / GPT-4	B. / GPT-4	B. / GPT-4	IOU	B. / GPT-4	B. / GPT-4	B. / GPT-4	Acc	IOU	B. / GPT-4	B. / GPT-4
- Caption-based Model												
mPLUG [36] (**4F**)	0.0	19.9 / 3.9	**25.7** / **6.0**	22.1 / **11.2**	0.0	14.9 / 3.0	**24.2** / **6.9**	**20.8** / **18.8**	0.0 / 0.0	0.0	19.7 / 4.0	**21.2** / **8.1**
GIT (L.M.) [61] (**4F**)	0.0	22.4 / 3.6	0.0 / 0.0	17.0 / 8.9	0.0	14.4 / 3.8	0.0 / 0.0	7.1 / 11.3	0.0 / 0.0	0.0	19.4 / 8.2	0.0 / 0.0
GIT (L.V.) [61] (**4F**)	0.0	**33.3** / **4.0**	0.0 / 0.0	**25.9** / 10.0	0.0	**20.5** / **4.2**	0.0 / 0.0	10.5 / 12.0	0.0 / 0.0	0.0	**29.8** / **8.6**	0.0 / 0.0
- Instruction-based Model												
VideoChat [37] (**8F**)	0.0	44.0 / 17.9	45.4 / 31.9	20.2 / 31.7	0.0	21.7 / 5.9	22.8 / **17.7**	7.3 / 31.1	67.5	0.0	47.4 / 8.2	43.1 / 44.6
Video-ChatGPT [47] (100**F**)	0.0	39.9 / **24.3**	40.1 / 24.9	36.5 / 41.2	0.0	**45.8** / 6.6	45.2 / 9.1	30.9 / **48.8**	**85.4**	0.0	**50.8** / **11.2**	43.3 / 40.4
mPLUG-Owl [60] (**4F**)	0.0	44.5 / 10.7	**47.3** / **35.0**	29.8 / **48.8**	0.0	43.0 / 5.0	**44.7** / 10.6	23.9 / 36.3	66.7	0.0	46.4 / 8.6	43.9 / 30.9
Video-LLaMA [74] (**8F**)	0.0	**48.4** / 7.7	42.9 / 29.0	46.5 / 34.1	0.0	45.5 / 7.2	41.1 / 17.2	42.3 / 31.2	64.2	0.0	50.1 / 10.2	39.0 / 28.0
LLaVA-NeXT [42] (64**F**)	0.0	47.9 / **41.3**	**49.5** / **69.8**	28.8 / **52.5**	0.0	**46.1** / **28.1**	**46.9** / **30.2**	26.9 / 43.8	48.2	0.0	48.7 / **55.0**	**44.9** / 38.3
VideoLLaVA [38] (64**F**)	0.0	39.0 / 9.7	42.1 / 30.3	24.7 / 35.0	0.0	37.0 / 9.1	37.0 / 13.25	20.8 / 36.3	67.2	0.0	43.3 / 36.7	36.8 / **54.0**
Otter (D.C.) [45] (128**F**)	0.0	30.2 / 7.7	32.3 / 28.3	21.7 / 20.0	0.0	28.7 / 1.7	32.9 / 7.9	17.7 / 36.3	45.0	0.0	32.5 / 2.1	27.3 / 36.8
Otter (FunQA) [35] (128**F**)	0.0	38.4 / 8.9	42.6 / 31.7	**47.5** / 32.1	0.0	40.0 / **7.3**	41.1 / 8.8	**44.5** / 38.8	69.4	0.0	44.7 / 10.3	**44.5** / **47.5**
Video-ChatGPT [47] + **FunMentor (Ours)**	0.0	**65.2** / **33.2**	**57.5** / 36.5	50.2 / **65.1**	0.0	**66.3** / **14.2**	**58.7** / **23.4**	45.3 / **52.2**	**85.4**	0.0	**55.1** / **13.3**	**46.3** / 54.8
Otter (FunQA) [35] + **FunMentor (Ours)**	0.0	33.4 / 13.4	37.8 / **45.8**	**58.3** / 34.2	0.0	60.4 / 11.0	44.4 / 9.3	**53.9** / 43.5	69.4	0.0	43.5 / 12.81	38.91 / **56.4**

our method to the HumorQA dataset. Overall, the performance of the models on the FunQA dataset is generally unsatisfactory. However, after fine-tuning VLM with FunMentor, a notable improvement is shown in its ability to comprehend counter-intuitiveness. We have made several key findings:

Timestamp localization task is the most challenging. Caption-based models focus mainly on captioning and often omit temporal information. Due to this, they are not equipped to deal with our task that emphasizes temporal aspects, so they are not scored. Similarly, Instruction-based models, such as Otter, take visual information from specific frames without temporal context. Their outputs are thus confined to individual frames, making them ineffective at addressing temporal localization. In summary, none of the current VLMs could solve H1, C1, or M1 tasks since they do not have a sense of time (refer to Appendix D.2). To enhance the timestamp localization task analysis, we add comparative baselines (as shown in Table 5): **Random Guess**, a specialized time-sensitive VLM (**TimeChat** [55]), and a novel method: clipping and rating (**CR**), previous Video Temporal Grounding (VTG) tasks have utilized this method (UniVTG [41]). We have also applied this approach to VideoQA. Results show that both UniVTG

Table 5: Timestamp Localization Task's baseline. The result uses IOU metrics.

Model	H1	C1	M1
Random	29.4	32.6	21.1
TimeChat	8.4	14.6	6.0
UniVTG (**CR**)	**59.6**	39.1	**37.9**
LLaVA-NeXT (**CR**, 64**F**)	58.0	**97.2**	33.6

Table 6: Human Performance on HumorQA. The average scores using BLEURT / GPT-4. Full results are in Appendix D.

Task	H2	H3	H4
Model SOTA	65.2 / 33.2	57.5 / 45.8	58.3 / 65.1
Human	70.0 / **83.6**	63.1 / **84.4**	67.3 / **79.1**

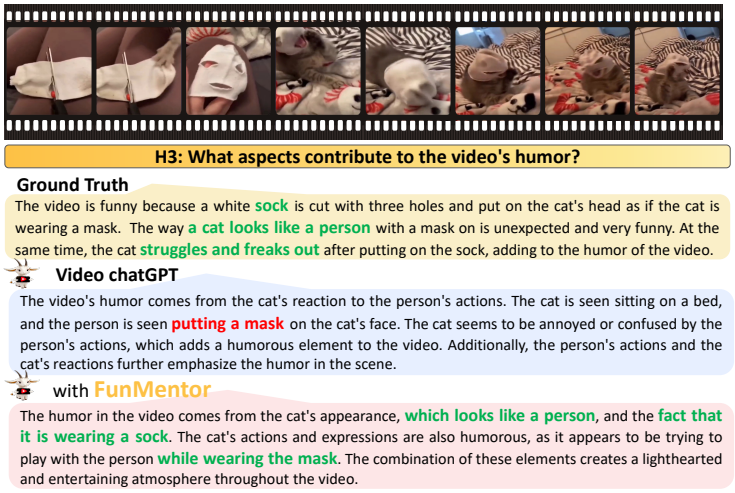

Fig. 4: VLM responses on before and after FunMentor. Here shows the answers given by Video-ChatGPT [47] on HumorQA video before and after FunMentor. From the GroundTruth, it is evident that **three** key elements contribute to the humor in the video: **"sock"**, **"mask"**, and **"cat resemble a human"**. Initially, Video-ChatGPT only identified the mask, failing to grasp the full essence of the video's humor. However, after the combination with FunMentor, Video-ChatGPT successfully recognized the **"looks like a person"** and the **"fact that it is wearing a sock"**, thus demonstrating a true understanding of what makes the video funny.

and VLMs employing the clipping and rating method have achieved significant improvement. multimodal language model.

No clear winner across all tasks. Caption-based models excel in providing detailed descriptions but struggle in tasks that require reasoning, resulting in a notable performance gap between description tasks (e.g., H2) and reasoning tasks (e.g., H3). On the other hand, instruction-based models demonstrate stronger reasoning capabilities than caption-based models but tend to underperform in description tasks. One possible explanation is that instruction-based models may generate excessive information in their answers, including a significant amount of incorrect information. We conducted experiments to compare machine versus human performance on FunQA as shown in Table 6.

Performance varies greatly across different video types. Generally CreativeQA is the most challenging especially in reasoning. For instance, the GPT-4 scores of Video-ChatGPT and Otter on C3-Reasoning are notably low. One possible reason is that humor and magic videos often depict daily life that models have encountered previously, whereas creative videos that models have never seen before, causing them unable to understand and generate reasonable answers.

Insufficient evaluation metrics for free-text tasks. Traditional metrics yield near-zero scores on free-text questions (refer to the complete FunQA Benchmark results in the Appendix D), as they solely focus on short textual similarity. While BLEURT scores are significantly higher, they still fall short in evaluating more complex similarities. Intuitively, GPT-4 is found to show preliminary capabilities in assessing free-text in deep understanding, which will be detailed

in Appendix D.1. However, there are still issues of instability, where the same content can receive different scores.

Finetuned Otter performs well yet with limitations. Otter (FunQA) shows obvious performance advantages over Otter (D.C.). As shown in Table 4, Otter (FunQA) performs better in BLEURT and GPT-4. However, there are still some limitations that Otter (FunQA) falls short of reaching SOTA scores. One possible reason revealed is that the input of Otter is only 128 frames sampled from the video, which is insufficient for comprehensive reasoning.

FunMentor demonstrates the powerful potential of agent-based fine-tuning Our proposed FunMentor outperforms the previous best method by 10.8 for the GPT-4 score of H3 task, with improvements of 4.8 and 20.9 for Otter (FunQA) and Video-ChatGPT, respectively. Additionally, the results reveal that FunMentor achieves significant performance improvements for Video-ChatGPT, particularly in the H2 and H4 tasks, while the improvement for H3 is relatively modest. This indicates that counter-intuitive reasoning remains a challenging aspect. The substantial performance enhancement by FunMentor highlights the promising prospects of agent-based fine-tuning methods. In the context of VLM requiring extensive training data, research in this direction holds the potential to uncover the vast capabilities of VLM.

5.4 Comparison with Previous Benchmarks

We chose NExT-QA as our comparative benchmark, which is designed to emphasize model reasoning abilities. NExT-QA also provides a multi-choice version of the dataset (NExT-OE), similar to FunQA-MC.

Table 7: Model's Performance in NExT-OE. This table illustrates that while the previous SOTA model, HGA, excels in the classic WUPS metric, it underperforms with GPT-4. See Appendix D.1 for a discussion on potential issues with WUPS.

Metrics	WUPS	GPT-4
HGA (SOTA on NExT-OE)	25.18	25.06
Otter (D.C.)	1.26	64.89
Otter (FunQA)	0.79	**73.06**

Traditional metrics are ineffective on VLMs' response. Table 7 shows the performance on NExT-OE of Otter and HGA [21] using different metrics. We evaluated these models using both traditional evaluation metrics (specifically, WUPS [43] that is employed by NExT-OE) and our novel evaluation metrics based on GPT-4. Otter exhibits a notably low performance on the WUPS metric that is drastically different than the GPT-4 score, primarily because WUPS is ill-suited for evaluating sentence-based responses and fares poorly when assessing phrases. Specific examples can be found in Appendix D.3.

NExT-QA is not a challenge for GPT-4. Our investigation into VLMs' performance on NExT-QA and FunQA datasets, shown in Table 8, reveals key insights. In the multi-choice format, GPT-4, even without video frames, scores comparably to VLMs on NExT-QA but resembles random guessing on FunQA-MC. Furthermore, while GPT-4V scores 80 and 61 on NExT-QA and FunQA-MC

Table 8: Performance of VLMs in NExT-QA and FunQA. N.QA, F.MC, and F.MC-R denote NExT-QA, FunQA-MC and FunQA-MC-R, which are multi-choice datasets and we use **accuracy** as the metric. Regarding N.QE (NExT-OE) and FunQA, which are open-ended answer datasets, we employ metrics based on **GPT-4**. * indicates that this model utilizes a Chat-box mechanism, and we employ GPT-3.5 to automatically convert its output into multiple-choice answers; † signifies the GPT-4 model without any visual information input; § denotes an sample version (10%).

Metrics	Acc			GPT-4	
Dataset	N.QA	F.MC	F.MC-R	N.QE	FunQA §
Random	20	20	20	-	-
Otter (D.C.) *	35	27	17	54	22
Otter (FunQA) *	42	31	26	58	**28**
GPT-4 †	44	34	23	30	2
GPT-4V(ision) (4**F**)	**80**	**61**	**39**	**79**	5

respectively, it drops to 39 on FunQA-MC-R, which focuses on counter-intuitive reasoning. This suggests that NExT-QA's inferential questions are no longer challenging for GPT-4, and the significant score difference in FunQA variants underlines GPT-4V's struggle with complex, non-intuitive questions.

FunQA is challenging for VLMs and GPT-4V. As Table 8 indicates, GPT-4 scores well on NExT-OE even without video access. The contrast in its performance on NExT-OE versus FunQA highlights FunQA's complexity in video reasoning. A significant performance disparity between GPT-4 and its video-enhanced version, GPT-4V, on FunQA again underscores the value of video content, corroborating findings from Table 3. Overall, FunQA stands out from prior benchmarks with its focus on reasoning skills and high-quality QA pairs deeply linked to video content, establishing itself as a robust benchmark in the LLM era for assessing VLMs' capabilities in counter-intuitive reasoning.

6 Limitations and Future Work

This paper has two limitations. **1)** The current FunQA dataset primarily contains video-level data and annotations. There is potential for enhanced video reasoning through denser annotations, akin to PVSG [67], which might include detailed spatial, temporal, and object-level annotations. **2)** The initial annotations were made in Chinese and later translated into English. While GPT was used to refine and complete the Chinese text, ensuring comprehensiveness and correctness, differences due to cultural differences between the languages might still persist. Looking ahead, we plan to enrich FunQA with more detailed and varied annotations. We aim to develop new metrics for a more accurate evaluation of models, particularly for open-ended questions. Our goal is to steer models towards deeper video reasoning. However, to ensure fair comparisons and prevent data leakage, it is advised that future research does not utilize the FunQA testing set.

Acknowledgments and Disclosure of Funding

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221- 0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Attardo, S.: A primer for the linguistics of humor. The primer of humor research **8**, 101–156 (2008) 2
2. Billig, M.: Laughter and ridicule: Towards a social critique of humour. Sage (2005) 2
3. Bitton-Guetta, N., Bitton, Y., Hessel, J., Schmidt, L., Elovici, Y., Stanovsky, G., Schwartz, R.: Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images (2023). https://doi.org/10.48550/ARXIV.2303.07274, https://arxiv.org/abs/2303.07274 5
4. Boyd, B.: Laughter and literature: A play theory of humor. Philosophy and literature **28**(1), 1–22 (2004) 20
5. Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., Poria, S.: Towards multimodal sarcasm detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Florence, Italy (7 2019) 3, 5
6. 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) 5, 21
7. Chiang, C.H., Lee, H.y.: Can large language models be an alternative to human evaluations? arXiv preprint arXiv:2305.01937 (2023) 10
8. Corporation, N.T.N.: Kasou taishou. https://www.ntv.co.jp/kasoh/index.html, [Accessed 23-Apr-2023] 20
9. Epstein, D., Chen, B., Vondrick, C.: Oops! predicting unintentional action in video. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 5
10. Fu, J., Ng, S.K., Jiang, Z., Liu, P.: Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166 (2023) 10
11. Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: Violet: End-to-end video-language transformers with masked visual-token modeling. arXiv preprint arXiv:2111.12681 (2021) 5
12. 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) 5, 21
13. Gao, D., Zhou, L., Ji, L., Zhu, L., Yang, Y., Shou, M.Z.: Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. arXiv preprint arXiv:2212.09522 (2022) 5
14. 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) 5, 21
15. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering (2017) 4

16. Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Agqa: A benchmark for compositional spatio-temporal reasoning (2021) 5, 21
17. Gruner, C.R.: Understanding laughter: The workings of wit & humor. Burnham Incorporated Pub (1978) 2
18. Hasan, M.K., Rahman, W., Bagher Zadeh, A., Zhong, J., Tanveer, M.I., Morency, L.P., Hoque, M.E.: UR-FUNNY: A multimodal language dataset for understanding humor. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2046–2056. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1211, https://aclanthology.org/D19-1211 3, 5
19. Hessel, J., Marasović, A., Hwang, J.D., Lee, L., Da, J., Zellers, R., Mankoff, R., Choi, Y.: Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest (2022) 5
20. Jang, Y., Song, Y., Kim, C.D., Yu, Y., Kim, Y., Kim, G.: Video Question Answering with Spatio-Temporal Reasoning. IJCV (2019) 3, 5, 21
21. Jiang, P., Han, Y.: Reasoning with heterogeneous graph alignment for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020) 13
22. Kamalloo, E., Dziri, N., Clarke, C.L., Rafiei, D.: Evaluating open-domain question answering in the era of large language models. arXiv preprint arXiv:2305.06984 (2023) 10
23. Kant, I.: Critique of judgment. Hackett Publishing (1987) 20
24. 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) 23
25. Koestler, A.: The act of creation. In: Brain Function, Volume IV: Brain Function and Learning, pp. 327–346. University of California Press (2020) 20
26. Kougia, V., Fetzel, S., Kirchmair, T., Çano, E., Baharlou, S.M., Sharifzadeh, S., Roth, B.: Memegraphs: Linking memes to knowledge graphs (2023) 5
27. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017) 10
28. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Li, F.F.: Visual genome: Connecting language and vision using crowdsourced dense image annotations (2016) 4
29. Lamont, P., Wiseman, R.: Magic in theory: An introduction to the theoretical and psychological elements of conjuring. Univ of Hertfordshire Press (2005) 20
30. Latta, R.L.: The basic humor process: A cognitive-shift theory and the case against incongruity. De Gruyter Mouton (1999) 20
31. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling–supplementary file 5
32. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: EMNLP (2018) 5, 21
33. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvqa+: Spatio-temporal grounding for video question answering. In: Tech Report, arXiv (2019) 3, 5, 21
34. Li*, B., Zhang*, P., Zhang*, K., Pu*, F., Du, X., Dong, Y., Liu, H., Zhang, Y., Zhang, G., Li, C., Liu, Z.: Lmms-eval: Accelerating the development of large multimoal models (March 2024), https://github.com/EvolvingLMMs-Lab/lmms-eval 3

35. 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) 5, 10, 11, 30
36. 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) 10, 11, 30
37. 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) 5, 10, 11, 30, 37
38. Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C.: Beyond rnns: Positional self-attention with co-attention for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8658–8665 (2019) 5
39. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection (2023), https://arxiv.org/abs/2311.10122 10, 11
40. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) 10
41. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding (2023) 11
42. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/ 10, 11
43. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems **27** (2014) 13
44. Martin, M.W.: Humour and aesthetic enjoyment of incongruities. The British Journal of Aesthetics **23**(1), 74–85 (1983) 2
45. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640 (2019) 3
46. 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) 20
47. Muhammad Maaz, Hanoona Rasheed, S.K., Khan, F.: Video-chatgpt. https://github.com/mbzuai-oryx/Video-ChatGPT (2023) 5, 10, 11, 12, 30
48. 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) 3, 5, 21
49. Nelms, H.: Magic and showmanship: A handbook for conjurers. Courier Corporation (2012) 20
50. Noordewier, M.K., Breugelmans, S.M.: On the valence of surprise. Cognition & emotion **27**(7), 1326–1334 (2013) 2
51. OpenAI: Gpt-4 technical report (2023) 5
52. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002) 10
53. Patro, B.N., Lunayach, M., Srivastava, D., Sarvesh, Singh, H., Namboodiri, V.P.: Multimodal humor dataset: Predicting laughter tracks for sitcoms. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 576–585 (January 2021) 3, 5

54. Pu, A., Chung, H.W., Parikh, A.P., Gehrmann, S., Sellam, T.: Learning compact metrics for mt. In: Proceedings of EMNLP (2021) 10
55. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. ArXiv **abs/2312.02051** (2023) 11
56. Runco, M.A., Jaeger, G.J.: The standard definition of creativity. CREATIVITY RESEARCH JOURNAL **24**(1), 92–96 (2012) 20
57. Seo, P.H., Nagrani, A., Schmid, C.: Look before you speak: Visually contextualized utterances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16877–16887 (2021) 5
58. 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) 3, 5, 21
59. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015) 10
60. Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., Qu, J., Zhou, J.: Is chatgpt a good nlg evaluator? a preliminary study. arXiv preprint arXiv:2303.04048 (2023) 10
61. 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) 10, 11, 30
62. 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 5, 21
63. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa:next phase of question-answering to explaining temporal actions (2021) 3, 5, 7, 21
64. 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) 3, 21
65. 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) 3, 21
66. 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) 23
67. Yang, J., Peng, W., Li, X., Guo, Z., Chen, L., Li, B., Ma, Z., Zhou, K., Zhang, W., Loy, C.C., et al.: Panoptic video scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18675–18685 (2023) 14
68. 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) 5, 21
69. 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) 10, 11, 30

70. 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 21
71. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning (2020) 3, 4, 5, 21
72. 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) 3, 21
73. 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) 3, 5, 21
74. 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 10, 11, 30
75. Zhao, Z., Zhang, Z., Xiao, S., Yu, Z., Yu, J., Cai, D., Wu, F., Zhuang, Y.: Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In: IJCAI. vol. 2, p. 8 (2018) 5
76. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018) 3
77. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images (2016) 4