



FunQA: Towards Surprising Video Comprehension

Binzhu Xie^{*,†}, Sicheng Zhang^{*,†}, Zitang Zhou^{*,†},
 Bo Li[†], Yuanhan Zhang[‡], Jack Hessel[§], Jingkang Yang[‡], Ziwei Liu[†][✉]

[†] Beijing University of Posts and Telecommunications, Beijing, China

[‡] S-Lab, Nanyang Technological University, Singapore

[§] The Allen Institute for AI, WA, USA

<https://github.com/Jingkang50/FunQA>

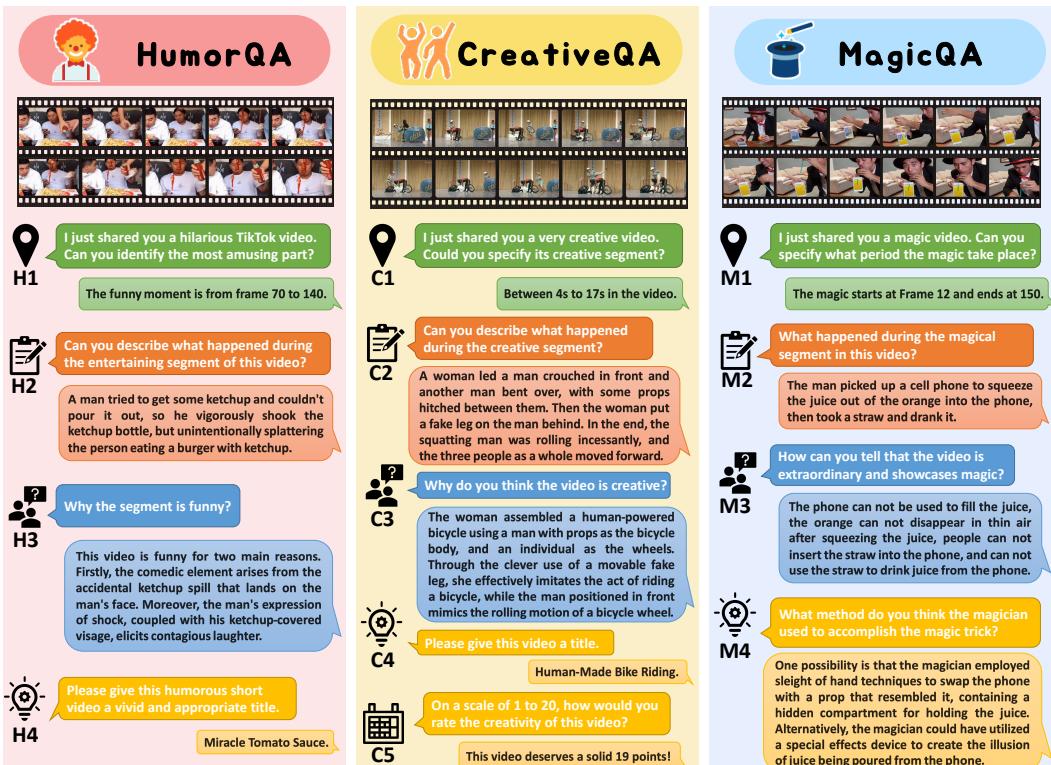


Figure 1: **Overview of FunQA.** FunQA comprises three subsets of surprising videos: 1) *HumorQA*, 2) *CreativeQA*, and 3) *MagicQA*. Each subset is associated with three common tasks: 1) *counter-intuitive timestamp localization*, 2) *detailed video description*, and 3) *reasoning around counter-intuitiveness* (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 the depth of video reasoning based on counter-intuitive and fun

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

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. Extensive experiments with existing VideoQA models 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, be they funny, creative, or filled with visual illusions, offer enjoyment and commands attention from viewers. This type of media elicits *positive surprise* [1],² 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 [2]. 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 Figure 1 (left) as an illustrative example. We witness a man engrossed in his phone, sharing a meal with friends. Suddenly, one of his companions squeezes a generous amount of ketchup, which, instead of adorning the fries, splatters onto the man’s face. The shock in his eyes, combined with his ketchup-covered visage, elicits laughter.³ 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 individuals were *gathered to enjoy a meal together*, and discern that the comedic element arises from the *ketchup intended for the fries ending up on the man’s face instead*, and that the innocent expression of the unfortunate victim indicates no significant harm was caused.

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. Examples of commonly employed VideoQA datasets include YouCook2 [6] which contains video clips from 2K cooking videos, Howto100M [7] which consists of only instructional videos. While there exist video datasets that explore the humor in TV shows [8, 9] and include tasks such as predicting laughter tracks [10], these tasks often heavily rely on audio and narrative cues, with visual clues might playing a lesser role.

To address this gap and evaluate computer vision models’ ability to identify and understand visual commonsense violations in videos, we introduce **FunQA**, a comprehensive and high-quality VideoQA dataset comprising 4.3K surprising videos and 312K manually annotated free-text QA pairs. Our dataset consists of three subsets: 1) *HumorQA*, 2) *CreativeQA*, and 3) *MagicQA*. Each subset covers different sources and video 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. 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

²c.f., *negative surprise*, e.g., a surprising medical bill.

³The hostility/superiority theory of humor posits that humor can arise from claiming superiority over someone or something [3, 4]; but alternate (more optimistic) theories of humor exist, [5] offers a survey.

Table 1: **Comparison between FunQA and other existing benchmarks.** Compare 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. **Anno.**: Annotation; **M**: Manual, **A**: Automatic; For Input, **V**, **A**, **S**, and **B** denote Video, Audio, Subtitle, and Bounding-box; **VC** means visual-centric, **Desc.** means Description, **Expl** for Explanation, **STR** for Spatial-temporal Reasoning. For QA Tasks, **MC** denotes Multiple Choice QA, **OE** means Open Ended QA, and **FT** means Free Text QA.

Datasets	Domain	Anno.	Video			Question Answer					
			Avg length (s)	# Clips (K)	Input	# QA pairs (K)	VC	Desc.	Expl.	STR	QA Task
MarioQA [11]	Games	M	3.6	188	V, A	188	Yes	No	Yes	Yes	OE
TGIF-QA [12]	Social Media	M	3.1	71.7	V	165.2	Yes	Yes	No	Yes	MC & OE
MovieQA [13]	Movies	A	202.7	6.77	V, A, S	6.4	No	No	Yes	Yes	MC
CLEVRER [14]	Synthetic Video	M	5	20	V	305	Yes	No	Yes	Yes	OE
TVQA [15]	TV shows	M	76.2	21.8	V, A	152.5	No	No	Yes	No	MC
TVQA+ [16]	TV shows	M	7.2	4.2	V, A, B	29.4	No	No	Yes	Yes	OE
Social-IQ [17]	Web videos	M	99	1.25	V, A, S, B	7.5	Yes	No	Yes	No	MC
NExT-QA [18]	Daily life	M	44	5.4	V, A	52	Yes	Yes	Yes	Yes	MC & OE
KnowIT-VQA [19]	TV shows	M	60	12	V, A, S	24	Yes	No	Yes	Yes	MC
AGQA [20]	Social Media	A	30	9.6	V, A	192	Yes	No	Yes	Yes	OE
AVQA [21]	Social Media	M	60	9.3	V, A	57.3	Yes	No	Yes	Yes	MC
STAR [22]	Daily life	A	-	22	V	60	Yes	No	No	Yes	MC
Env-QA [23]	Egocentric Video	M	20	23.3	V	85.1	Yes	No	No	Yes	MC
FIBER [24]	Daily life	M	10	28	V, A, S	2	No	Yes	No	Yes	OE
HumorQA	Daily life	M	7	1.8	V, A	141.3	Yes	Yes	Yes	Yes	OE & FT
CreativeQA	Performance	M	48	0.9	V, A	78.7	Yes	Yes	Yes	Yes	OE & FT
MagicQA	Magic	M	10	1.6	V, A	91.9	Yes	Yes	Yes	Yes	OE & FT
FunQA	Surprising videos	M	19	4.3	V, A	312	Yes	Yes	Yes	Yes	OE & FT

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. Additionally, we propose auxiliary tasks that pose higher-level challenges within the benchmark including assigning an appropriate and vivid title to the video, etc. 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 and discernment.
- 3) **Comprehensive Evaluation:** We have done an extensive and 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 [25, 26, 27], 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 [14]. To address the VideoQA task, the research community has introduced various benchmarks. As depicted in Table 1, Most commonly used VideoQA datasets are sourced from human-centric videos like movies [13], TV shows [15, 16, 19], and social media [12, 17, 18, 20, 21, 22, 24], and there are also object-centric datasets of game videos [11], synthetic videos [14] and egocentric videos [23]. MovieQA [13] and TVQA [15] 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 [12] 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 [18], 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 Solutions Earlier studies have explored various models, including LSTMs and graph-based neural networks, to capture cross-modal information [28, 29]. With the advent of Transformers, video understanding models like ClipBERT [30] and CoMVT [31] emerged, focusing on the comprehension of specific frames within a video. Subsequent models, such as Violet [32], extended their ability to encompass temporal and spatial information. However, these methods have primarily been applied to short videos. In the realm of long videos, MIST [33] stands out by achieving state-of-the-art performance and excelling in terms of computation efficiency and interpretability. Furthermore, recent Vision Language Models (VLMs) such as [34, 35, 36] 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 [37], which emphasizes weird, unusual, and uncanny images, and MemeGraphs [38], which revolves around memes featuring humor and sarcasm. Furthermore, some work even challenges models to comprehend complex multimodal humor in comics [39]. In the realm of large vision-language models, exemplified by GPT-4 [40], there is a particular focus on showcasing their ability to provide explanations for funny pictures. However, when it comes to videos, existing datasets exploring humor in TV shows or comedy tend to heavily rely on audio and narrative cues [8, 9, 10], with visual clues playing a comparatively lesser role.

3 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 in Figure 2. We introduce our construction pipeline in the end.

3.1 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 [41, 42] between reality and expectations, flourishing with the skillful juxtaposition and transformation of events [43, 44, 45]. 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 [46]. This program, showcasing original and novel skits performed by various amateur groups and judged by a panel, boasts a strong creative flair [47]. 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.

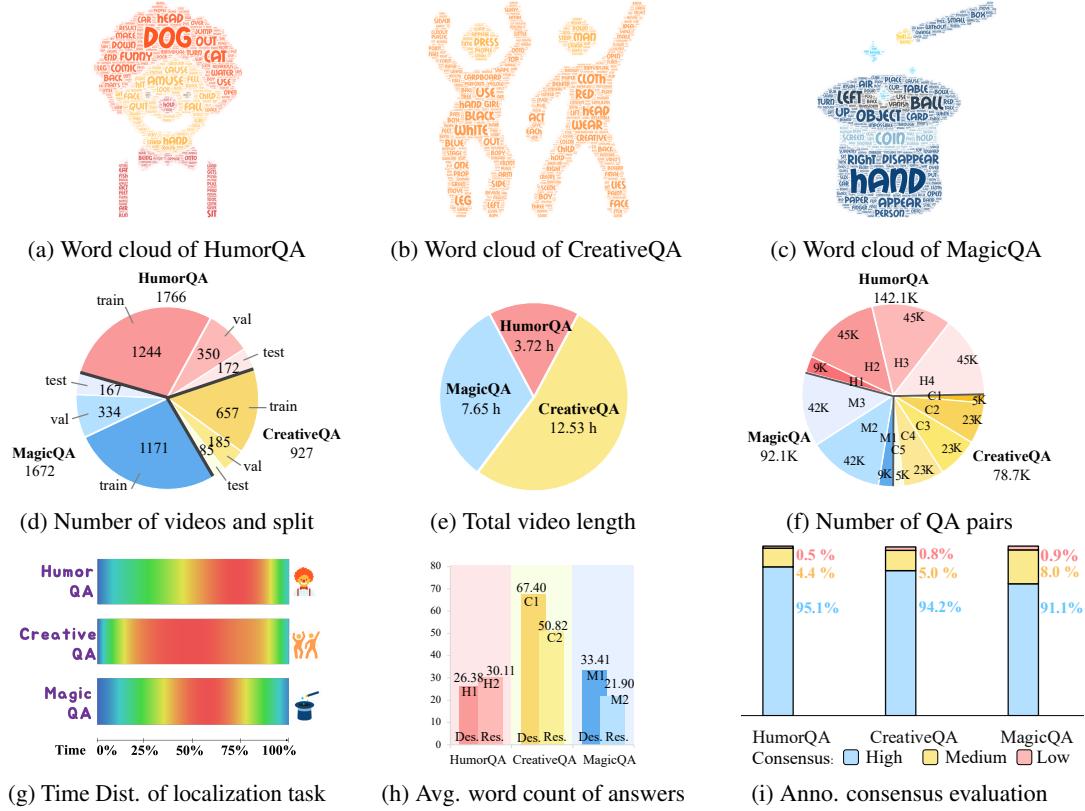


Figure 2: Statistics of FunQA Dataset. FunQA consists of three subsets, each corresponding to different video types, and is annotated with free-text QA pairs. The first row displays word clouds representing critical annotations for each subset. The second row provides key dataset statistics, including the number of videos for different splits, video length, and QA pair count for three subsets. In the last row, (g) highlights the high-frequency time span of the answer for localization questions in red, (h) shows the average word count of answers, and (i) presents the percentage of consensus between annotators for the same answer in a sampled set.

MagicQA MagicQA encapsulates 1672 magic performance videos sourced from across the web, spanning various genres like camera magic, close-up magic, and stage magic. The essence of magic revolves around the creation of seemingly impossible illusions [48], 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 [49] of these transformations.

3.2 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. Across all three subsets, 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 0 and 20. *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. In addition, some visual QA benchmarks [37] have begun to adopt free-text answers; this flexible answer format enables multimodal LLMs to go beyond multiple choice options (Large Language Models) [35, 36]; for Video QA, however, free-text are less common. To bridge this gap, we also consider some **free-text answer** evaluations, demanding enhanced video comprehension capabilities from the model. It will also assess the model’s ability to generate reasonably lengthy textual responses.

3.3 Dataset Statistics

FunQA contains **4,365** counter-intuitive video clips and **311,950** question-answer pairs, the total length of these videos is **23.9h** and the average length of video clips is **19** seconds. FunQA consists of three fine-grained subsets, each one containing well-designed tasks. The specific numbers and split of videos can be seen in Figure 2 (d). Each subset’s video lengths can be seen in Figure 2 (d). The specific number of QA pairs for each task can be seen in Figure 2 (e).

For our localization task, the timestamp heat map for the three different types of videos can be seen in Figure 2 (g), 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., 8.7 in Activity-QA [50], 11.6 in NExT-QA [18]). The specific word count of each task is shown in Figure 2 (h). FunQA has a well-established annotation process and high annotation quality, the result of our annotation consensus evaluation can be seen in Figure 2 (i). For each video category, more than 90% of the annotations exhibit a high level of consensus, with only 1% of the content showing low consensus. Approximately 8% of the data demonstrates variations in consensus, thus highlighting the objectivity of our dataset.

HumorQA, CreativeQA, and MagicQA word clouds are shown in Figure 2 (a-c). More statistics and FunQA full word cloud are given in Appendix A.3.

3.4 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, and we paid crowd-workers a target of ¥25/hr. More details of the dataset construction pipeline can be seen in Appendix A.1.

Pre-processing Initially, we crawled videos from TikTok, Bilibili, and YouTube (these videos are a collection of surprising videos). Then we performed a two-stage manual cleaning and cutting process on the collected videos to ensure counter-intuitive features and video quality and to exclude non-ethical content and sensitive information, resulting in a cut video clip.

Manual Annotation We annotated the videos according to the characteristics of different tasks design 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 of annotation, we conducted a secondary annotation of 10% of the tasks and performed Consensus Evaluation to ensure the objectivity of our annotations.

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 increased the number of QA pairs to 312K. Additionally, we created diverse task types, such as FunQA multiple-choice and FunQA dialogue. More details are given in Appendix A.2 and Appendix B.

4 Experiments

In this section, we present an introduction to caption-based and instruction-based models, followed by an exploration of diverse metrics for evaluating FunQA tasks. Our comprehensive experiments and deep analysis of the results are then presented. More details are given in Appendix C.

4.1 Baselines

4.1.1 Caption-based Models

mPLUG mPLUG [51] consists of two unimodal encoders for image and text independently, a cross-modal skip-connected network, and a decoder for text generation. Based on the connected representation of the image and prefix sub-sequence, the decoder is trained with a prefix language modeling loss by generating the remaining caption.

GIT GIT [52] is composed of one image encoder and one text decoder, for videos, multiple frames are sampled and encoded independently, and features are added with an extra learnable temporal embedding before concatenation. The image encoder is based on the contrastive pre-trained model [53] with raw image input and a compact 2D feature map output. The text decoder is a transformer module to predict the text description.

4.1.2 Instruction-based Models

VideoChat VideoChat [35] is an end-to-end chat-centric video understanding system, the version we use is VideoChat-13B. Its VideoChat-Embed architecture is instantiated using BLIP-2 [54] and StableVicuna (13b-delta), and combines pre-trained ViT-G [55] and GMHRA [56]. For the token interface, VideoChat-Embed uses the pre-trained QFormer with an additional linear projection to output Video Embedding. Both video description and Video Embedding will be input in LLMs (LLAMA-13B [57]) for multimodal understanding and output timestamped video text descriptions.

Video-ChatGPT Video-ChatGPT [36] is a large vision-language model with a dedicated video-encoder and LLM, which feeds the video frames into pre-trained video encoder, adds spatio-temporal features and feeds them into linear layer. Video Embedding is then input into LLM (Vicuna-7B, v1.1) along with System command and User Query to output the answer. It uses a data-centric, human-assisted, and semi-automated annotation framework for high-quality video instructional data with unique multimodal (visual-verbal) capabilities.

Otter The Otter model [34] utilizes the OpenFlamingo training paradigm, where the pretrained OpenFlamingo model consists of a LLaMA-7B language encoder [57] and a CLIP ViT-L/14 vision encoder [58]. In the fine-tuning process for instruction tuning, Otter freezes both the encoders and only fine-tunes the Perceiver resampler module. We evaluate two versions of Otter: one that is fine-tuned on the Dense Caption dataset [59], and another that is fine-tuned on the FunQA training set. Due to GPU memory limitations, Otter can only train and test on 128 frames from a video.

4.2 Evaluation Metrics

Timestamp Localization (H1, C1, M1) We employ the intersection of union based on time span.

Description & Reasoning (H2-4, C2-4, M2-3) For all the **free-text** tasks, we employ three approaches for evaluation. Firstly, we utilize traditional NLG(Natural Language Generation) metrics. We use BLEU-4 [60], ROUGE-L [61], CIDEr [62], and BLEURT [63] as our metrics. The first two rely on N-gram overlap, which is only sensitive to lexical variations and cannot identify changes in

Table 2: **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 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 **Difference** 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		
	H1	H2	H3	H4	C1	C2	C3	C4	C5	M1	M2	M3
- Caption-based Model												
mPLUG [47]	-	1.5 / 16.4 / 1.0 19.9 / 16.0	1.1 / 12.5 / 0.4 25.7 / 18.1	0.6 / 7.5 / 0.1 22.1 / 17.3	-	0.4 / 13.4 / 0.0 14.9 / 24.3	0.7 / 12.6 / 0.1 24.2 / 9.0	0.3 / 3.2 / 0.0 20.8 / 13.7	-	1.2 / 15.8 / 0.5 19.7 / 16.9	0.9 / 8.9 / 0.4 21.2 / 8.8	-
GIT (L.M.) [52]	-	0.5 / 12.8 / 0.2 22.4 / 22.0	-	1.1 / 7.7 / 0.7 17.0 / 26.8	-	0.0 / 6.4 / 0.0 14.4 / 5.0	-	0.3 / 1.5 / 0.2 7.1 / 25.2	-	0.2 / 11.2 / 0.1 19.4 / 12.7	-	-
GIT (L.V.) [52]	-	1.2 / 16.9 / 0.6 33.3 / 31.5	-	1.0 / 8.8 / 0.7 25.9 / 33.2	-	0.1 / 8.3 / 0.0 20.5 / 5.0	-	0.5 / 2.8 / 0.4 10.5 / 23.3	-	0.6 / 13.7 / 0.1 29.8 / 21.4	-	-
- Instruction-based Model												
VideoChat [35]	-	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	-
Video-ChatGPT [36]	-	0.5 / 14.0 / 0.1 39.9 / 20.7	0.7 / 12.4 / 0.1 40.1 / 33.0	0.4 / 3.2 / 0.2 18.6 / 47.5	-	1.1 / 19.8 / 0.2 45.8 / 19.1	0.8 / 17.3 / 0.1 45.2 / 30.1	0.2 / 1.9 / 0.2 18.8 / 44.5	85.4	0.7 / 20.8 / 0.0 50.0 / 11.8	0.5 / 11.3 / 0.0 43.3 / 29.2	-
Otter (D.C.) [34]	-	1.1 / 14.3 / 0.4 30.2 / 9.8	1.2 / 14.2 / 0.4 32.3 / 13.9	0.5 / 5.4 / 0.1 21.7 / 13.3	-	0.5 / 13.8 / 0.1 28.7 / 11.0	1.0 / 16.8 / 0.2 32.9 / 10.6	0.3 / 2.3 / 0.1 17.7 / 4.2	45.0	1.0 / 15.0 / 0.3 32.5 / 14.4	1.1 / 12.8 / 0.2 27.3 / 13.7	-
Otter (FunQA) [34]	-	1.5 / 18.1 / 0.9 38.4 / 22.2	1.3 / 15.4 / 0.5 42.6 / 31.0	0.8 / 5.9 / 0.5 24.5 / 51.3	-	1.5 / 19.6 / 0.5 40.0 / 21.9	2.2 / 21.2 / 0.5 41.1 / 21.1	0.3 / 4.3 / 0.3 21.7 / 43.9	69.4	2.6 / 23.8 / 1.6 44.7 / 18.4	3.4 / 20.3 / 2.6 44.5 / 19.8	-

sentence semantics or grammar. The latter two are reference-based evaluation metrics. Secondly, several works [64, 65, 66, 67] 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 C.1.

Creative Scoring (C5) The evaluation uses the formula: $Metrics = 100 \times \left(1 - \frac{|Predict - GT|}{20}\right)$.

4.3 Results and Observations

Our results are summarized in Table 2. As an illustration, the responses of different models on HumorQA can be seen in Figure 3. Overall, the performance of the models on the FunQA dataset is generally unsatisfactory, and we have made several key findings:

Timestamp localization task is the most challenging. Caption-based models, due to their emphasis on captioning tasks, tend to provide descriptions of the entire video even when tasked with timestamp localization (refer to Appendix C.2). Conversely, instruction-based models, which are typically derived from image-based VLMs, focus on specific keyframes rather than considering the entire temporal space of the video.

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 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.

Performance varies greatly across different video types. Most models can get a relatively accurate answer in humor and magic videos but have difficulty answering the questions in creative videos. One possible reason is that humor and magic videos often depict daily life that models have encountered previously, whereas creative videos contain content that models have never seen before, rendering them unable to generate new ideas and resulting in irrelevant and erroneous answers.

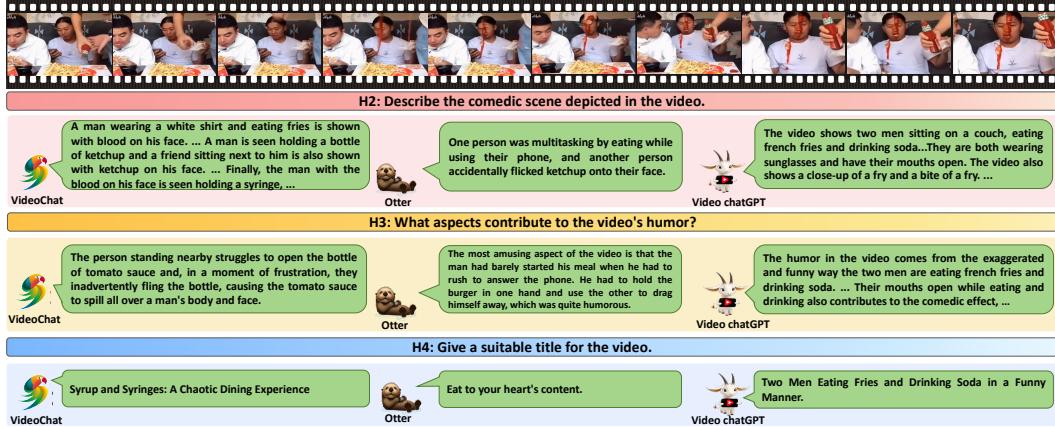


Figure 3: **Model responses on a HumorQA example.** Here shows the answers given by VideoChat [35], Video-ChatGPT [36], and Otter [34] on HumorQA video in Figure 1. On task **H2, H3**, VideoChat has the best performance. On task **H4**, Video-ChatGPT and Otter answer better, which is in line with our experiment result in Table 2. However, the answers from all models are still far from the ground truth. The descriptions of details and counter-intuitive explanations have numerous shortcomings. For example, Video-ChatGPT added incorrect details to the description, such as "wearing sunglasses", the humorous reason for "throwing ketchup" was wrongly interpreted by VideoChat as "knocking over the ketchup bottle", etc.

Insufficient evaluation metrics for free-text tasks. Traditional metrics yield near-zero scores on free-text questions, 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 C.1. However, there are still issues of instability, where the same content can receive different scores.

Finetuned Otter performs well on traditional metrics but lags behind in GPT-4 score. We finetuned Otter on Dense Caption and FunQA, and Otter (FunQA) shows obvious performance advantages over Otter (D.C.). While Otter performs better in traditional metrics like ROUGE-L compared with other instruction-based models, the GPT-4 score of Otter (FunQA) underperforms. One possible reason revealed is that the input of Otter is only 128 frames sampled from the video, which is insufficient for comprehensive reasoning. Besides, the discrepancy between Otter’s scores on traditional metrics and GPT-4 matches our finding of the lack of evaluation metrics.

5 Limitations and Future Work

This paper has two limitations. **1)** Current FunQA dataset mainly includes video-level data and annotations, but denser annotations can be developed to explore more possibilities of video reasoning. Examples include detailed spatial and temporal annotations, such as captions corresponding with specific time axes and annotations of object level. **2)** The raw annotations are completed by the annotator in Chinese. In the process of translating into English, we first use GPT to polish and supplement Chinese annotations, making the text as thorough as possible, as short or incomplete text may result in misunderstanding. However, there may still be differences due to cultural diversities between the two languages.

In the future, we expect to expand our dataset with denser and more diverse annotations. Also, new metrics will be explored to better evaluate models’ performance, especially in open-ended questions which lack in-depth metrics. Finally, we hope to provide directions for advancing models toward deeper video reasoning. Specific ideas include that Otter can sample more frames from the input video to improve causal reasoning capacity.

Acknowledgments and Disclosure of Funding

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221- 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] Marret K Noordewier and Seger M Breugelmans. On the valence of surprise. *Cognition & emotion*, 27(7):1326–1334, 2013. [2](#)
- [2] Mike W Martin. Humour and aesthetic enjoyment of incongruities. *The British Journal of Aesthetics*, 23(1):74–85, 1983. [2](#)
- [3] Charles R Gruner. *Understanding laughter: The workings of wit & humor*. Burnham Incorporated Pub, 1978. [2](#)
- [4] Michael Billig. *Laughter and ridicule: Towards a social critique of humour*. Sage, 2005. [2](#)
- [5] Salvatore Attardo. A primer for the linguistics of humor. *The primer of humor research*, 8:101–156, 2008. [2](#)
- [6] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [7] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. [2](#)
- [8] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy, 7 2019. Association for Computational Linguistics. [2, 4](#)
- [9] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 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)*, pages 2046–2056, Hong Kong, China, November 2019. Association for Computational Linguistics. [2, 4](#)
- [10] Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh, Hunar Singh, and Vinay P. Namboodiri. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 576–585, January 2021. [2, 4](#)
- [11] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875, 2017. [3](#)
- [12] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video Question Answering with Spatio-Temporal Reasoning. *IJCV*, 2019. [3](#)
- [13] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. [3](#)

- [14] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020. 3
- [15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 3
- [16] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*, 2019. 3
- [17] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019. 3
- [18] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. 3, 4, 6
- [19] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020. 3
- [20] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning, 2021. 3
- [21] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491, 2022. 3
- [22] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 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. 3
- [23] Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. pages 1675–1685, October 2021. 3
- [24] Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan C. Stroud, and Rada Mihalcea. Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework, 2022. 3
- [25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 3
- [26] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images, 2016. 3
- [27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 3
- [28] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019. 4
- [29] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *IJCAI*, volume 2, page 8, 2018. 4
- [30] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling—supplementary file. 4
- [31] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 4

- [32] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 4
- [33] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. *arXiv preprint arXiv:2212.09522*, 2022. 4
- [34] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 4, 7, 8, 9
- [35] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 4, 6, 7, 8, 9
- [36] Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. Video-chatgpt. <https://github.com/mbzuai-oryx/Video-ChatGPT>, 2023. 4, 6, 7, 8, 9
- [37] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images, 2023. 4, 6
- [38] Vasiliki Kougia, Simon Fetzel, Thomas Kirchmair, Erion Çano, Sina Moayed Baharlou, Sahand Sharifzadeh, and Benjamin Roth. Memographs: Linking memes to knowledge graphs, 2023. 4
- [39] Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest, 2022. 4
- [40] OpenAI. Gpt-4 technical report, 2023. 4
- [41] Immanuel Kant. *Critique of judgment*. Hackett Publishing, 1987. 4
- [42] John Morreall. Philosophy of Humor. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition, 2023. 4
- [43] Robert L Latta. *The basic humor process: A cognitive-shift theory and the case against incongruity*. De Gruyter Mouton, 1999. 4
- [44] Brian Boyd. Laughter and literature: A play theory of humor. *Philosophy and literature*, 28(1):1–22, 2004. 4
- [45] Arthur Koestler. The act of creation. In *Brain Function, Volume IV: Brain Function and Learning*, pages 327–346. University of California Press, 2020. 4
- [46] Nippon Television Network Corporation. Kasou taishou. <https://www.ntv.co.jp/kasoh/index.html>. [Accessed 23-Apr-2023]. 4
- [47] Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *CREATIVITY RESEARCH JOURNAL*, 24(1):92–96, 2012. 4, 8
- [48] Henning Nelms. *Magic and showmanship: A handbook for conjurers*. Courier Corporation, 2012. 5
- [49] Peter Lamont and Richard Wiseman. *Magic in theory: An introduction to the theoretical and psychological elements of conjuring*. Univ of Hertfordshire Press, 2005. 5
- [50] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueling Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 6
- [51] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections, 2022. 7

- [52] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 7, 8
- [53] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyu Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021. 7
- [54] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 7
- [55] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. 7
- [56] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning, 2022. 7
- [57] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 7
- [58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 7
- [59] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 7
- [60] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 7
- [61] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 7
- [62] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 7
- [63] Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for mt. In *Proceedings of EMNLP*, 2021. 7
- [64] Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*, 2023. 8
- [65] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023. 8
- [66] Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*, 2023. 8
- [67] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023. 8

A More details of FunQA dataset

A.1 Dataset Construction Pipeline

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 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

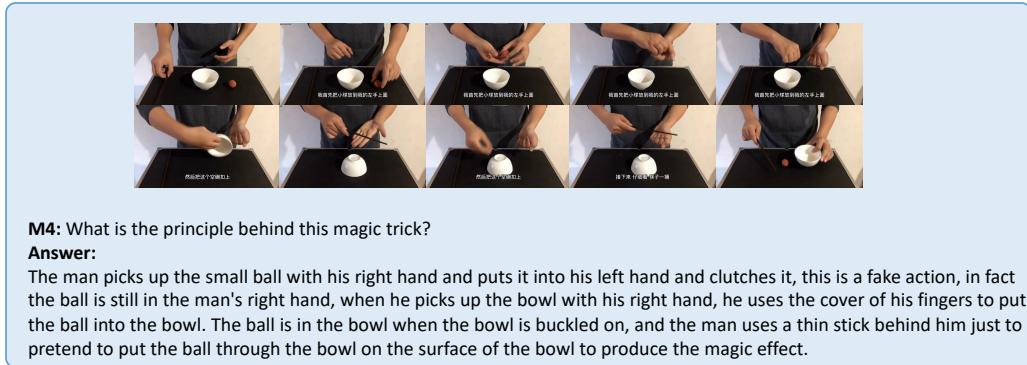


Figure 4: **Example of Magic Method Task.** During the annotation process, we discovered that even as humans, it is difficult to fully understand the complete principles behind the implementation of magic tricks right from the beginning.

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]

Figure 5: **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.

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 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. Figure 4 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

Short description of H2.				
Annotation Content	Task	Task requirement	Reason for incorrection	Modifications
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	"Prepared 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.

Figure 6: **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.

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.

Figure 7: **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.

video clip, etc) can be seen in Figure 5. Figure 6 also illustrates some errors encountered during the annotation process and the suggested modifications provided. 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. For content downloaded from streaming platforms, we obtained permission from the respective uploaders. For *Kasou Taisho*, we submitted a fair use request to the official website of Tokyo TV in Japan.

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. Figure 7 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 Figure 8.

A humor video with its translated Description and Explanation

Video ID: H_H_114_3686_3881.mp4

Description: A woman was sitting inside a shopping cart and was pushed by another person back to the shopping cart return area. The cart was eventually locked, and the person even waved goodbye.

Explanation: The video is funny because most people don't sit inside a shopping cart, so it feels novel, and in the end, the person is locked inside the cart and returned to the designated area. The woman inside the shopping cart is in a precarious situation, which is amusing.



User Input

购车人

It's a great Chinese title that mentions the two key things in the video, the shopping cart and the people, and it comes from the Chinese word for 购物车 (shopping cart) and changes the word for 物 (things) to 人 (people).

Ambiguous GPT output

CH2Eng Translation: Human trafficking
 Expansion_with_Eng_1: Human smuggling
 Expansion_with_Eng_2: Human smuggling
 Expansion_with_Eng_3: Trafficking in persons
 Expansion_with_Eng_4: Human smuggling
 Expansion_with_Eng_5: Human trade

Manual modification

Expansion_with_Eng_1: Shopping cart with a person inside
 Expansion_with_Eng_2: Person found in the shopping cart
 Expansion_with_Eng_3: Inhabited shopping cart
 Expansion_with_Eng_4: Occupied shopping cart
 Expansion_with_Eng_5: Shopping cart, but there's someone inside

Figure 8: **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.

A.3 Data Statistics

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 3. The word cloud shaped by FunQA logo of the all annotation word is shown in Figure 9.

Table 3: Statistics of the FunQA raw data.

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



Figure 9: Word cloud of FunQA free-text answer.

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

B.1 FunQA Multi-choice Dataset

FunQA 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 Figure 10.

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

Figure 10: **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.

B.2 FunQA Dialog Dataset

Most of the current LLMs are in the form of dialogues. To cater to their data input, we produced the FunQA 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 Figure 11.

C More Details of Experiment

C.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 annotated video as a reference to score the new title. Each Prompt can be seen in Figure 12, 13, and 14.

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

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.

Figure 11: **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.

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.

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

GPT Prompt for Title 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.

Figure 12: **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.

traditional evaluation metrics (e.g. synonym substitution, utterance rearrangement) giving reasonable evaluation.

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. Figure 15 shows an example of GPT-4 generating different responses to the same content.

C.2 On Failure of the Timestamp Localization Task

There is no model that has the ability to answer the Timestamp Localization Task. Figure 16 shows exactly what is answered for each model.

C.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_MSRTTT, which were fine-tuned on the video captioning task for the VATEX and MSRTTT datasets, respectively.

VideoChat VideoChat, specifically the VideoChat-13B version, is an end-to-end system for video

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.

Figure 13: **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.

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.

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.

Figure 14: **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.

All of our experiments took five to six days on a single NVIDIA A40 48GB GPU.

C.4 More Examples

Figure 17 shows the responses of different models on CreativeQA and MagicQA.

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:

- Language Expression (Fluency, Vocabulary):** The text is fluent and uses appropriate vocabulary, providing a coherent scene description. Score: **5/5**
- 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**
- Common Sense Errors:** The text doesn't contain any obvious common-sense errors; it describes a plausible scene. Score: **10/10**
- 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**
- 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**
- 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:

- 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**
- 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**
- Common Sense Errors:** There are no obvious common-sense errors. Score: **10/10**
- 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**
- Details:** Text2 does not provide any of the details mentioned in Text1. It seems to describe a completely different scenario. Score: **0/35**
- 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**.

Figure 15: **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.

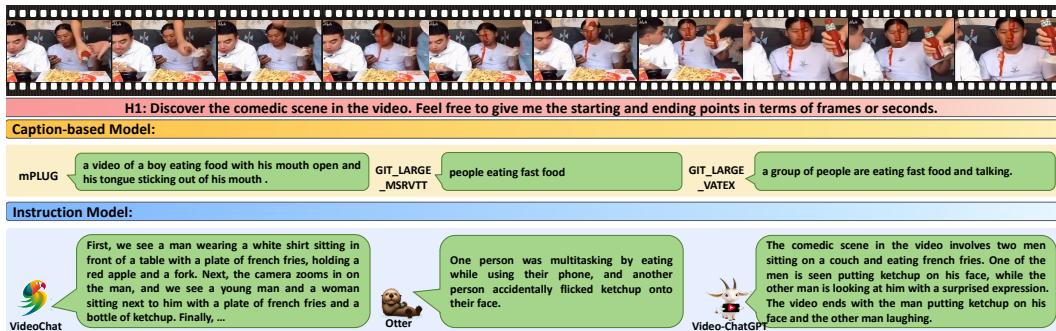


Figure 16: **Model responses on Timestamp Localization Task.** As can be seen, the model gives answers to the Timestamp Localization Task still focusing on describing the video content and still does not answer the specific time period when asked to type the number of frames and seconds.



C2: Provide a detailed account of the video's imaginative scene.

VideoChat

First, we see a man standing on a stage in a theater wearing a mask. Next, we see a bicycle on stage with a sign that says "cycle racing". Then, we see a woman wearing a skirt and a jacket on stage, followed by a man wearing a hat. Then, we see a bike on stage with a sign that says "bike show"..... Overall, the video gives a glimpse into a world of bicycle racing shows and competitions, showcasing the unique art of human and bicycle synchronization.

Otter

A person pushed a large blue cardboard, from which a person in black clothes with red stripes on one side emerged.... A man took out a bicycle-shaped cardboard and placed it in front of him. After that, the man put the cardboard back in place, flipped the cardboard to reveal the rear side, and revealed the red circle and the words "repair shop" written on it.

Video-ChatGPT

The video shows a man dressed in a red and white costume, standing on a stage and holding a microphone. He is wearing a clown mask and appears to be a comedian. The stage is set with a red curtain and a large screen behind him. The man is talking to the audience and making jokes, while the screen displays a picture of a clown. The man then starts to juggle a red ball, and the audience laughs.

C3: Why is the video as a whole creative?

VideoChat

...The various signs and people dressed up in different costumes ... This video highlights the artistry and entertainment of bicycle racing events, aiming for a fun and enjoyable experience for viewers. The bicycle serves as a canvas for various designs and performances, such as a giant bicycle with a person inside it, people wearing costumes, and even a woman dressed as a bicycle....

Otter

The creative point of this video lies in using a person dressed in a bicycle costume to simulate the scene of riding a bike. The handlebar is managed by a string to make the person shake their body to simulate the dynamic process of cycling. Another innovative aspect is the use of cardboard to mimic the wheels of the bicycles.

Video-ChatGPT

The video is creative because it combines elements of humor, juggling, and a clown-themed performance. The man dressed in a clown costume and holding a microphone delivers jokes and engages the audience, while the red curtain and the large screen behind him create a theatrical atmosphere.

C4: Suggest a title that captures the creativity in the video.

VideoChat

Human-Bicycle Synchronization in a World of Performance and Creativity.

Bicycle.

Otter

Video-ChatGPT

The Clown's Comedy and Juggling Show.



M2: Give a narration of the video's magical segment.

VideoChat

In the beginning of the video, we see a young man wearing a black suit and a hat sitting in a room, holding a soda cup in his hand. He then stands up and starts to move around the room. The scene then cuts to a man in a black suit, sitting in a chair while holding a soda cup. He also shows us the smartphone in his hand again....Finally, we see the man sitting on a chair, holding his smartphone in his hands, and we see a black table in the background. The video ends with the man still holding his phone.

Otter

By swiping his hand across the screen of his mobile phone, a man triggers a transformation of chess pieces from virtual to tangible.

Video-ChatGPT

The video features a man in a suit playing a game on his cell phone. He is seen holding the phone in his hand and tapping the screen. The video also shows a man in a suit playing a game on a tablet. The man in the suit is seen holding the tablet in his hand and tapping the screen. The video ends with the man in the suit holding the tablet and tapping the screen.

M3: Why is the video as a whole magic?

VideoChat

The video is considered magical because it shows a young man using his smartphone in various ways, such as playing games, making strange figures, and showing off his phone to the viewer. The man's actions with his phone seem unrealistic and mysterious, making the video feel like it is capturing some kind of magic....Overall, the combination of the man's actions with his phone and the changing scenes creates a sense of magic that captures the viewer's attention.

Otter

It is impossible for objects to levitate, and the contents displayed on a screen cannot transcend the screen and enter the physical world.

Video-ChatGPT

The video is considered magic because it showcases the man in a suit playing a game on a cell phone and a tablet, seemingly at the same time. This is a remarkable feat, as it is not possible for a person to hold and play with two devices simultaneously. The video captures the man's ability to multitask and engage with both devices simultaneously, which is a remarkable and impressive display of skill.

Figure 17: **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.