v-HUB: A Benchmark for Video Humor Understanding from Vision and Sound

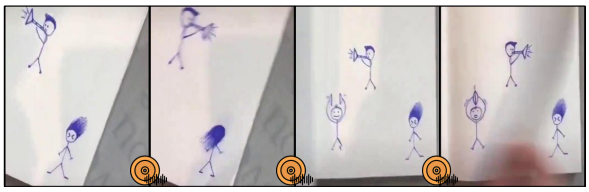
Zhengpeng Shi[1,3] **Yanpeng Zhao**[3] † ✉ **Jianqun Zhou**[2,3] **Yuxuan Wang**[4]
Qinrong Cui[4] **Wei Bi**[4] **Songchun Zhu**[3] **Bo Zhao**[1] ✉ **Zilong Zheng**[3] ✉
[1]Shanghai Jiao Tong University [2]Wuhan University
[3]State Key Laboratory of General Artificial Intelligence, BIGAI
[4]Independent Researcher

Abstract

AI models capable of comprehending humor hold real-world promise—for example, enhancing engagement in human-machine interactions. To gauge and diagnose the capacity of multimodal large language models (MLLMs) for humor understanding, we introduce v-HUB, a novel video humor understanding benchmark. v-HUB comprises a curated collection of non-verbal short videos, reflecting real-world scenarios where humor can be appreciated purely through visual cues. We pair each video clip with rich annotations to support a variety of evaluation tasks and analyses, including a novel study of environmental sound that can enhance humor. To broaden its applicability, we construct an open-ended QA task, making v-HUB readily integrable into existing video understanding task suites. We evaluate a diverse set of MLLMs, from specialized Video-LLMs to versatile OmniLLMs that can natively process audio, covering both open-source and proprietary domains. The experimental results expose the difficulties MLLMs face in comprehending humor from visual cues alone. Our findings also demonstrate that incorporating audio helps with video humor understanding, highlighting the promise of integrating richer modalities for complex video understanding tasks.

1 Introduction

Humor enriches our daily lives and appears in many forms, from jokes and cartoons to comedies and viral videos. AI models capable of understanding humor hold promise for engaging with humans empathetically (Hampes, 2001, 2010), but perceiving and comprehending humor can be challenging even to humans due to the heavy reliance on nontrivial reasoning, social and cultural contexts, etc (see Figure 1). This, on the other hand, makes humor understanding a promising testbed to evaluate how well state-of-the-art AI models understand humor. Indeed, there has been a line of research centering around gauging the capability of pre-trained large language models (LLMs) for humor understanding from visuals and text (Hessel et al., 2022; Hyun et al., 2024; Ko et al., 2023; Chen et al., 2024), but parallel studies of *multimodal* LLMs are limited, though they are naturally suited for understanding multimodal humor.



(a) **Visuals**+Audio. As the man flips through the pages, cartoon characters gradually appear, accompanied by a distinct melody. First, the dancer's rhythm and the suona player's piercing tune, then the cymbal player's resonant clash, together creating an evolving effect.

(b) **Visuals**+Audio+Text. A guy messaged his friend that he was making a birthday cake for them. After it was baked and sliced, the inside mimicked their chat bubble layout. The whole scene was made merrier by the Happy Birthday melody.

Figure 1: Examples of visual-centric humor understanding, where 'audio' and 'text' refer to environmental sound (*cf.* human speech) and visual text, respectively.

In this work, we address this gap by investigating humor understanding with multimodal LLMs (MLLMs), focusing specifically on MLLMs that can process video. We choose video as the primary medium of humor, since it captures nuanced variations and diverse styles, presenting a unique challenge. For example, perceiving the humor in

†: Project Lead. ✉: Corresponding Author.
Contact: shi_zpeng@sjtu.edu.cn,yannzhao.ed@gmail.com.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38544–38567

Table 1: Humor understanding benchmark comparison. Modality indicates the modalities used for humor understanding, where T, V, and A denote text, visual, and audio, respectively. Our v-HUB is *visual-centric*, as it contains humor derived predominantly from visual cues (V) and enhanced by environmental sound (A).

Dataset	Type	Source	Average Duration	Modality	Tasks		
					Explanation	Matching	Open-ended QA
NYCC (Hessel et al., 2022)	Cartoon	New Yorker Caption Contests	—	T, V	✓	✓	✗
Oogiri-GO (Zhong et al., 2024)	Images	Oogiri Game	—	T, V	✗	✓	✗
MUStARD (Castro et al., 2019)	Sitcom	TV Shows	19 sec.	T, V, A	✗	✓	✗
WITS (Kumar et al., 2022)	Sitcom	TV Shows	17 sec.	T, V, A	✗	✓	✗
UR-FUNNY (Hasan et al., 2019)	Speech	TED Talks	20 sec.	T, V, A	✗	✓	✗
SMILE (Hyun et al., 2024)	Sitcom, Speech	TV Shows, TED Talks	28 sec.	T, V, A	✓	✗	✗
ExFunTube (Ko et al., 2023)	Short videos	Youtube	16 sec.	T, V, A	✓	✗	✗
HumorQA (Xie et al., 2024)	Surprising videos	Youtube	7 sec.	T, V, A	✓	✓	✓
v-HUB (ours)	Short videos, **Silent comedies**	X, Youtube, Charlie Chaplin's **silent films**	**14** sec.	V, A	✓	✓	✓

Figure1b requires recognizing visual text (or background music) and the layout of chat bubbles and understanding their temporal and semantic correspondences with the cut surface of the cake slice.

While there have been a few benchmarks on video humor understanding (see Table 1), most of them are designed exclusively for the evaluation of LLMs (Ko et al., 2023; Hyun et al., 2024),[1] and the curated humor is dominated by linguistic cues, ignoring the fact that humans can understand humor from visual cues alone, exemplified by silent comedies. Though Xie et al. (2024) have conducted a humor understanding evaluation of MLLMs, the evaluation is limited in coverage and temporal complexity. Moreover, as an important component of non-verbal humor, environmental sound has been overlooked in their data curation and evaluation.

To address these limitations, we curate a set of visual-centric humorous videos from two complementary sources: Charlie Chaplin's silent films and user-generated short funny videos. Silent film humor is conveyed through visual cues, but is thematically and culturally constrained due to the scripted performance. To increase diversity, we incorporate user-generated funny short videos from various occasions and cultural backgrounds. We rigorously filtered the videos to retain only those where the humor has no reliance on speech and has a duration longer than 5 seconds. Our final dataset consists of videos where humor is derived predominantly from the visual modality, making it visual-centric and better suited for evaluating MLLMs, including variants that can natively process audio.[2]

To assess how well MLLMs understand humor in video, we create v-HUB, a video humor understanding benchmark. v-HUB offers two typical evaluation tasks for humor understanding. (1) First, the *Caption Matching* task challenges MLLMs to align video captions with the corresponding videos. (2) Second, the *Humor Explanation* task evaluates whether MLLMs can extract humor elements and provide accurate rationales. To broaden the applicability of v-HUB, we further construct (3) an *Open-ended QA* task that evaluates the MLLMs' fundamental understanding of videos from the humor genre across temporal, descriptive, and causal dimensions. Together, these tasks provide a comprehensive framework to benchmark MLLMs in visual-centric humor understanding.

We evaluate representative MLLMs from both open- and closed-source domains. Depending on the input modalities, we consider the following three task settings. (1) The *Text-Only* setting assumes human-level interpretation of video contents and provides detailed human-written descriptions. (2) The *Video-Only* setting offers only videos (without audio) to assess the ability of MLLMs to derive humor solely from visual cues. (3) We further propose a novel *Video+Audio* setting that combines visual and auditory signals to determine whether sound cues—such as background music and sound effects—help MLLMs (aka. OmniLLMs) better understand humor.

We empirically find that MLLMs generally perform better with text-only inputs than with video-only inputs. For example, Qwen2.5-VL drops in accuracy from 0.726 to 0.666 on *Caption Matching*, indicating its struggles in capturing subtle visual cues for humor understanding. Adding audio yields notable improvements across most OmniLLMs. For instance, MiniCPM2.6-o improves from 0.362 to 0.442 in accuracy on *Caption Matching*, though it still lags behind the text-only setting. Overall,

[1]They translated videos into language descriptions and performed verbal humor evaluation with LLMs.

[2]In this work, audio primarily refers to environmental sound rather than human speech (see Section 2.2).

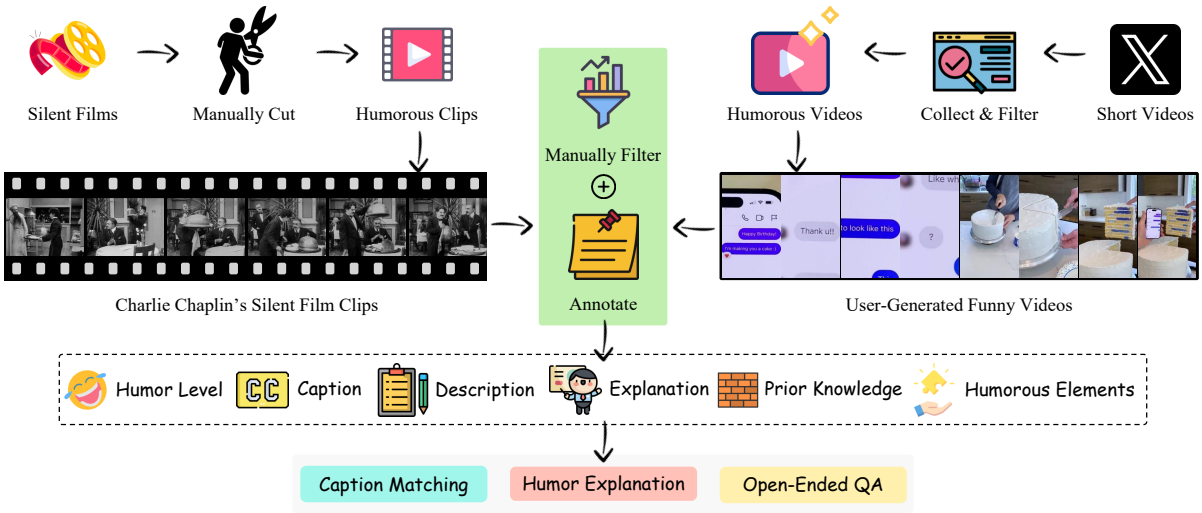


Figure 2: Data Curation Pipeline. To collect visual-centric humorous videos, the pipeline consists of two main stages: (a) *Humorous video collection*, where annotators identify timestamps of self-contained humorous clips for silent films and verify humor presence in short videos (see Section 2.1). (b) *Filtering and annotation*, where only visual-dominant humor is retained and annotated (see Section 2.3). The annotation is further used for task construction (see Section 3).

v-HUB presents a new challenge and contributes to a comprehensive evaluation of MLLMs. It exposes their weakness in visual-centric humor understanding, stresses the need for enhancing their visual reasoning capabilities, and highlights the promise of integrating additional modalities like sound for video understanding.

2 Visual-Centric Video Humor Curation

2.1 Humor Video Sources

Our goal is to collect humorous videos that are visual-centric and illustrate diverse humor. A straightforward approach is to collect humorous clips from silent comedies that are entirely devoid of speech. Though silent films may contain recorded music, sound effects, and few captions, which may contribute to the expression of humor, the humor primarily arises from the visual modality. A major issue with silent film clips is that they have rather narrow themes and employ limited storytelling techniques. To enhance the diversity of humor in our dataset, we further incorporate user-generated short funny videos from the Internet. Specifically, we select videos primarily from an X account (@humansnocontext) that frequently shares humorous clips with minimal reliance on speech or text-based context.[3] We also use a subset of YouTube videos from Xie et al. (2024) to diversify video sources. Thus, our dataset comprises humorous videos from two different domains that complement each other (see Figure 2):

- **Charlie Chaplin's Silent Films:** We reviewed Charlie Chaplin's classic silent films from 1914 to 1938 and collected 729 funny clips. Each humor is ensured to be self-contained, without relying on additional video contexts. Figure 2 shows an example in this domain.

- **User-Generated Funny Videos:** We reviewed the X user @humansnocontext's tweets posted between March 28, 2023 and October 12, 2024 and collected 18080 short funny videos. To ensure diverse video sources, we collected 1769 surprising YouTube videos from Xie et al. (2024).

2.2 Preprocessing and Filtering

We preprocess and filter the initially collected videos according to duration, appropriateness, and speech reliance, sequentially. (1) **Duration:** We retain videos ranging from 5 to 60 seconds long. Short clips under 5 seconds generally fail to convey meaningful humor, while clips exceeding 1 minute often rely on dialogue. For silent films, we segment long scenes to isolate individual humorous moments, ensuring that each segment captures the full humor, without becoming too long for generation tasks. (2) **Appropriateness:** To ensure that the contents of our videos are appropriate, we adhered to the safety objectives outlined in Thoppilan et al.

[3]Though most videos are sourced from a single X account, the content is sufficiently diverse (see Figure 4), as the account owner collects humorous videos from across the Internet.

Humor Level: Very Humorous

Humorous Elements: 'Visuals'; 'Visual Text'; 'Sound'

Descriptive Caption: Have you seen a message cake?

Text/Subtitle Presence: Visual text exists

Creative Caption: He even made a red heart.

Prior Knowledge: 'Happy Birthday' is playing.

Description: In the video, the guy told someone on the phone that he would give them a cake exactly like "this", and the listener was confused. After the cake was baked and cut, the cut surface revealed the same pattern as the chat bubble layout in the conversation.

Explanation: When the guy said the cake was exactly the same as "this", the listener didn't understand what he meant. But after the cake was cut, its cut surface revealed the same pattern as the chat bubble layout, which viewers found very humorous.

Figure 3: Example annotation of a short video that conveys humor through visuals, visual text, and background sound. Knowing the Happy Birthday melody makes the video merrier (see Section 2.3).

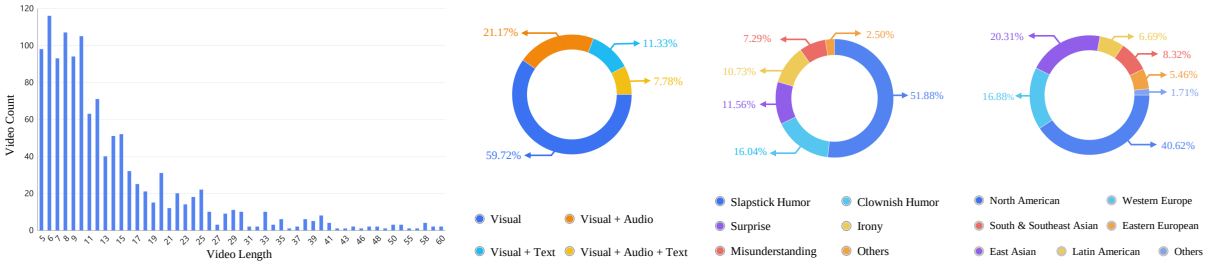
(2022) and excluded videos that violated the established criteria (see details in Appendix A.1). (3) **Speech Reliance:** We minimize reliance on speech. Since there is little to no speech in Charlie Chaplin's silent films, we primarily focused on user-generated funny videos and employed both manual and automatic approaches to filter out speech-heavy videos (see details in Appendix A.1).

2.3 Annotation

We recruited eight annotators based on the following criteria: (1) sufficient English proficiency to understand video content, (2) broad cultural knowledge to interpret humor arising from various contexts, and (3) strong observational skills assessed through a qualification test (see Appendix A.2). To ensure consistency, we provided detailed guidelines for each annotation task and created a reference manual for on-demand use. Each video underwent three rounds of annotation to guarantee correctness and thoroughness. We conducted the following primary annotation tasks (see Figure 3 for an example annotation):

- **Humor Evaluation:** Annotators independently evaluated whether the video was humorous.
- **Captioning:** Each annotator was asked to write two types of captions for each video, without seeing existing annotations, including captions and descriptions, from other annotators, thus ensuring an independent and unbiased judgment.
  - *Descriptive Captions* directly describe or highlight the humor present in the video content from the original publisher's perspective.
  - *Creative Captions* extend beyond the video's original humor by adding imaginative or novel elements (see the visual caption in Figure 5b).

  The dual-caption annotation supports a comprehensive assessment of humor in video from both comprehension and generation perspectives.
- **Video Description:** Annotators were instructed to describe the events in each video, including all the details necessary for understanding the humor, without making inferences, focusing only on observable objects, actions, and expressions. After the first annotator completes the video description, subsequent annotators review and refine the current descriptions for correctness and completeness.
- **Video Labeling:** Annotators labeled the key humor sources (e.g., human actions, objects, visual effects, or sound cues) in each video and noted whether any visual text was present. If an element appeared, but did not contribute to humor, it was not selected.
- **Background Knowledge:** Annotators determine whether understanding the humor in a video requires background knowledge, which refers to external contextual information that cannot be directly inferred from the videos but is necessary or helpful for understanding humor.
- **Humor Explanation:** Three annotators sequentially create and refine humor explanations by adding missing details, guaranteeing comprehensive coverage of the labeled humor sources



(a) (b) (c) (d)

Figure 4: Data statistics: (a) Distribution of video lengths; (b) Distribution of visual-centric groups; (c) Distribution of humor types, where *Others* contains Parody, Satire, Miscellaneous, etc; (d) Distribution of cultural backgrounds, where *Others* covers Middle Eastern and North African, etc.

through an iterative refinement process. In our annotation manual (see Figure 8 in Appendix) for humor explanation, we require that annotators include all the humorous elements they can find, as thoroughly as possible, and explain why these elements make people feel humorous.

2.4 Data Analysis

Duration. All videos in our final dataset are restricted to a duration of 5–60 seconds, with the majority concentrated within 30 seconds (see Figure 4a). This design ensures that humor is self-contained, sufficiently nuanced, and compatible with the context length limits of most MLLMs.

Diversity. To show v-HUB contains a variety of humor, we follow Buijzen and Valkenburg (2004) to categorize videos into five categories by humor type (see Figure 4c), and follow House et al. (2004) and Ronen and Shenkar (2013) to categorize videos into six categories by cultural background (see Figure 4d). The illustrations demonstrate that v-HUB covers a wide range of humor types and cultural backgrounds, thus it is sufficiently diverse.

Visual-centric. After all filtering processes, we were left with 1218 videos, including 267 silent humor clips from Charlie Chaplin and 951 user-generated short funny videos from the Internet. The total duration of the videos is 4.7h, and the average duration is around 14s. All of them rely on the visual modality to express humor. We identify two key modalities that dominate the delivery of humor: visuals and audio. Apart from 722 videos (59%) conveying humor primarily via pure visual cues (denoted by 'Visual'), 137 videos (11%) contain additional linguistic cues in visual form—such as embedded captions and subtitles (denoted by 'Visual+Text')—that extend humor, 256 video humor (21%) is enhanced by additional sound that covers non-speech auditory elements, such as background music and sound effects(denoted by 'Visual+Audio'), and 94 videos (8%) convey humor through visuals, sound, and visual text (denoted by 'Visual+Audio+Text'). The video distribution over the four groups is illustrated in Figure 4b.

Consensus Evaluation. For annotations like humor explanation and video description, the second and third annotators reviewed and modified previous annotations to ensure consistency. We employ Krippendorff's alpha (Krippendorff, 2011) to assess the annotators' consensus on the humor evaluation, using 'Low,' 'Medium,' and 'High' to indicate the strength of the consensus. In v-HUB, more than 90% of data demonstrated a 'High' consensus, while only 0.3% showed a 'Low' consensus.

3 v-HUB: A Visual-Centric Video Humor Understanding Benchmark

3.1 Evaluation Tasks

To comprehensively evaluate the capability of MLLMs in humor understanding, we propose three tasks that reflect different aspects of humor reasoning: Caption Matching, Humor Explanation, and Open-ended QA:

- **Caption Matching.** In this discriminative task, models must correctly associate videos with their corresponding captions. Unlike ordinary caption matching tasks, our design challenges MLLMs to go beyond surface-level matching and assess their ability to understand video humor that is pronounced by *creative captions* from a generation perspective. For each video with a creative caption, we randomly sample four *descriptive captions* from other videos as the distractors.

- **Humor Explanation.** In this generative task,

models must identify humor points within each video, provide coherent explanations, and reference relevant visual or auditory cues.

- **Open-ended QA.** To further assess the fundamental understanding of video content, we generate a set of open-ended question-answer pairs for each video (see details in Appendix B.1). These questions—automatically generated by GPT-4o and manually verified—encompass temporal, descriptive, and causal aspects (Xiao et al., 2021).[4] This extends the benchmark beyond humor-specific reasoning, providing a broader assessment of video reasoning skills.

3.2 Evaluation Methods

We employ different evaluation strategies depending on the task type:

- **Accuracy.** For *Caption Matching*, we measure accuracy to determine whether the model correctly identifies the most appropriate response.
- **Quality of Open-ended Responses.** For *Humor Explanation* and *Open-ended QA*, we adopt both automatic and human evaluation approaches:
  - *Semantic Similarity.* We compute similarity scores between model-generated answers and human-provided answers using BERTScore (Zhang* et al., 2020), which captures fine-grained semantic similarity beyond simple word overlap. In addition, we employ Sent-BERT (Reimers and Gurevych, 2019) to assess sentence-level semantic coherence, as well as METEOR (Banerjee and Lavie, 2005), which provides a more nuanced assessment of semantic adequacy and fluency.
  - AutoDQ (Wang et al., 2024a): This method evaluates the presence of humor-related events in the generated explanations. AutoDQ extracts key events from the model's output and compares them to ground truth (GT) annotations using entailment analysis. It provides three metrics: recall, precision, and F1 score (see Appendix B.2 for details). Unless otherwise specified, we report F1 scores.
  - *Human Evaluation.* We randomly sample a subset of model-generated explanations and compare them with ground truth. The evaluators rate the explanations based on accuracy and logicality, providing insight into the gap between human and MLLMs' explanations. We present and discuss the results in Table 9 in Appendix C due to limited space.

[4]There are 81, 742, and 395 QA pairs for temporal, descriptive, and causal questions, respectively.

4 Experiments

4.1 Experimental Setup

MLLMs. We consider both proprietary MLLMs, such as Gemini-2.5-Flash (Team et al., 2025) and GPT-4o (Hurst et al., 2024), as well as public MLLMs like Qwen2.5-VL (72B) (Bai et al., 2025), and Intern3.5-VL (8B) (Wang et al., 2025). OmniLLMs such as Video-SALMONN-2 (7B) (Tang et al., 2025), MiniCPM2.6-o (8B) (Yao et al., 2024), and Qwen2.5-Omni (7B) (Xu et al., 2025), which can process audio, are also included (an overview of all evaluated MLLMs is presented in Table 7).

Evaluation Settings. To understand the roles of different modalities in video humor understanding, we consider the following three settings: Text-Only, Video-Only, and Video+Audio, which means models are tested with text, video (w/ audio), and video-audio inputs, respectively.

- *Text-Only.* In this setting, models receive detailed human-written video descriptions; no visual or audio information is available to the models. Thus, it evaluates the language reasoning ability of MLLMs in isolation.
- *Video-Only.* Models are provided with only raw video frames (w/o audio). This setting assesses their intrinsic visual comprehension capabilities. Depending on the presence of visual text, we further divide results into two groups: 'w/ visual text' and 'w/o visual text'.
- *Video+Audio.* Models receive both video frames and audio signals, allowing us to examine whether the inclusion of auditory information improves humor understanding. Depending on the contribution of audio to humor, we further divide results into two groups: 'w/ humor audio' and 'w/o humor video'.

4.2 Main Results

Based on the results in Table 2, we analyze the humor competence of MLLMs along three dimensions: video humor discovery, understanding, and subtle humor inference. Our results reveal several shortcomings of MLLMs: they (i) struggle to

Table 2: Model performance on three tasks.

MLLMs	Explanation			Matching	Open-ended QA	
	SentBERT	METEOR	AutoDQ	Accuracy	SentBERT	METEOR
Text-Only						
Gemini-2.5-Flash	0.547	0.249	0.342	0.615	0.728	0.642
Video-SALMONN-2	0.571	0.246	0.317	0.359	0.595	0.435
MiniCPM2.6-o	0.546	0.236	0.325	0.531	0.562	0.454
Qwen2.5-Omni	0.536	0.233	0.316	0.656	0.719	0.546
Qwen2.5-VL	0.543	0.250	0.342	0.726	0.760	0.598
Intern3.5-VL	0.556	0.256	0.348	0.643	0.701	0.593
GPT-4o	0.560	0.255	0.374	0.762	0.690	0.645
Video-Only						
Gemini-2.5-Flash	0.459	0.199	0.175	0.580	0.424	0.270
video-SALMONN-2	0.269	0.150	0.052	0.243	0.317	0.169
MiniCPM2.6-o	0.381	0.165	0.112	0.362	0.369	0.186
Qwen2.5-Omni	0.384	0.159	0.144	0.553	0.382	0.121
Qwen2.5-VL	0.441	0.187	0.150	0.666	0.445	0.202
Intern3.5-VL	0.422	0.180	0.125	0.609	0.385	0.235
GPT-4o	0.455	0.192	0.206	0.646	0.411	0.286
Video+Audio						
Gemini-2.5-Flash	0.460	0.199	0.173	0.581	0.416	0.268
video-SALMONN-2	0.281	0.170	0.066	0.240	0.323	0.185
MiniCPM2.6-o	0.408	0.173	0.120	0.442	0.380	0.278
Qwen2.5-Omni	0.428	0.174	0.125	0.617	0.424	0.168

identify humorous elements when explicit cues are absent, (ii) inadequately fuse information across modalities for understanding, and (iii) show limited capacity for inferring subtle humor.

Limited ability in humor discovery. Across settings, MLLMs tend to perform better on open-ended QA than on humor explanation. This performance disparity reveals that they are limited in perceiving humor. For example, in the Text-Only setting, Qwen2.5-VL, whose SentBERT score drops from 0.760 in QA to 0.543 in humor explanation. These findings suggest that MLLMs are more successful when the question itself provides explicit cues that direct attention to a specific humorous element in the scene. By contrast, the humor explanation task, which requires models to independently identify and articulate the source of humor without such guidance, poses a greater challenge. This indicates that while MLLMs are often able to reason about humor once it is highlighted for them, they struggle with the more cognitively-demanding task of discovering humor directly from contextual cues.

Heavy reliance on linguistic cues for humor understanding. Comparing text-based and video-based evaluations, we observe marked differences across all three tasks, where the Text-Only setting yields substantially higher scores than the video-based settings, implying that current MLLMs are heavily dependent on linguistic cues for humor understanding. For example, on open-ended QA, Qwen2.5-VL achieves a SentBERT score of 0.760 with text input, but it plummets to 0.445 when presented with raw video (w/o audio). While the addition of audio provides a marginal but consistent performance boost, this gain is minimal compared to the contribution of text. This wide performance gap suggests that MLLMs' cross-modal fusion capabilities are still underdeveloped, leading them to rely on linguistic cues rather than effectively integrating visual and auditory signals.

Table 3: The impact of visual text on video humor understanding in the Video+Audio setting.

Models	Sound contributing to humor			Sound not contributing to humor		
	Matching	Open-ended QA		Matching	Open-ended QA	
	Accuracy	SentBERT	METEOR	Accuracy	SentBERT	METEOR
w/ visual text						
Gemini-2.5-Flash	0.621	0.440	0.303	0.715	0.437	0.288
video-SALMONN-2	0.200	0.336	0.202	0.255	0.332	0.193
MiniCPM2.6-o	0.453	0.466	0.357	0.511	0.430	0.320
Qwen2.5-Omni	0.716	0.442	0.174	0.686	0.434	0.171
w/o visual text						
Gemini-2.5-Flash	0.523	0.396	0.257	0.569	0.416	0.265
video-SALMONN-2	0.235	0.301	0.174	0.245	0.327	0.185
MiniCPM2.6-o	0.378	0.369	0.280	0.449	0.338	0.215
Qwen2.5-Omni	0.547	0.432	0.151	0.612	0.415	0.171

Incapability for subtle humor inference. The caption matching task goes beyond surface-level linking between literal descriptions and videos; instead, it requires models to find the *creative caption* that enhances or extends humor in the video. We find that most models exhibit limited performance (e.g., below 0.8), suggesting their incompetence for subtle humor inference. The difficulty is magnified when models process raw video. For example, video-SALMONN-2's accuracy falls from 0.359 in the Text-Only setting to 0.240 in the Video+Audio condition. This pronounced struggle to connect creative text to original visual humor context reveals a critical weakness in the models' capacity for the implicit cross-modal reasoning that is fundamental to comprehending sophisticated humor.

4.3 Further Analysis

To conduct a deeper analysis of model results, we further divide our experimental results based on previously annotated humor modalities and background knowledge essential for delivering humor in video, to analyze how different types of humor affect models' explanatory capability. We also presented supplementary quantitative results in Appendix C, such as human preference comparison of humor explanations across four model categories. (see Table 9).

Visual text adds value regardless of the audio's comedic contribution. As shown in Table 3, MLLMs generally perform better on videos containing visual text than on those without linguistic cues in the Video+Audio setting, except for video-SALMONN-2 on the caption matching

Table 4: The impact of background knowledge on video humor understanding in the Video+Audio setting.

MLLMs	Explanation			Matching	Open-ended QA	
	SentBERT	METEOR	AutoDQ	Accuracy	SentBERT	METEOR
w/ Background Knowledge						
video-SALMONN-2	0.467	0.173	0.114	0.324	0.396	0.198
MiniCPM2.6-o	0.515	0.203	0.193	0.447	0.427	0.204
Qwen2.5-Omni	0.512	0.199	0.176	0.663	0.493	0.219
w/o Background Knowledge						
video-SALMONN-2	0.285	0.177	0.025	0.252	0.301	0.174
MiniCPM2.6-o	0.440	0.180	0.115	0.417	0.358	0.259
Qwen2.5-Omni	0.459	0.181	0.127	0.615	0.441	0.157

task when sound contributes to humor. For example, Gemini-2.5-Flash achieves SentBERT and METEOR scores of 0.440 and 0.303 for open-ended QA with visual text, compared to 0.396 and 0.257 without it. When sound does not contribute to humor, the benefit of visual text remains evident: Gemini-2.5-Flash improves from 0.416 to 0.437 in open-ended QA SentBERT and from 0.569 to 0.715 in matching accuracy with visual text. These results indicate that *visual text serves as a useful complementary cue for humor understanding*, particularly when audio signals provide limited informative content.

Knowledge-based cues facilitate humor understanding. We identified 374 videos that require contextual background knowledge and evaluated MLLMs under two settings: with and without the explicit provision of such knowledge. As shown in Table 4, MLLMs consistently achieve higher performance when background knowledge is provided under the Video+Audio setting. For instance, Qwen2.5-Omni attains a SentBERT and AutoDQ scores of 0.512 and 0.176 on the humor explanation with background knowledge, compared to 0.459 and 0.127 without. These findings suggest that while MLLMs implicitly encode multiple cultural contexts, their *comprehension of humor is significantly enhanced by the explicit provision of background knowledge*, underscoring the central role of linguistic and knowledge-based cues in complex video humor understanding tasks.

MLLMs have greater difficulty in comprehending humor in historically distant videos. We analyze the performance of MLLMs under the Video-Only setting across two subsets from distinct eras: last-century Charlie Chaplin's silent films (CCSF) and contemporary user-generated funny videos (UGFV). As shown in Table 5, MLLMs consistently achieve higher scores on UGFV across all evaluation metrics. For example, Gemini-2.5-Flash attains a SentBERT of 0.469 for humor explanation

Table 5: The impact of video era on video humor understanding in the Video-Only setting.

MLLMs	Explanation			Matching	Open-ended QA	
	SentBERT	METEOR	AutoDQ	Accuracy	SentBERT	METEOR
Last-Century *Charlie Chaplin's Silent Films*						
Gemini-2.5-Flash	0.422	0.188	0.130	0.562	0.386	0.221
video-SALMONN-2	0.281	0.146	0.012	0.165	0.296	0.154
MiniCPM2.6-o	0.343	0.150	0.097	0.307	0.314	0.128
Qwen2.5-Omni	0.339	0.144	0.096	0.494	0.337	0.119
Contemporary *User-Generated Funny Video*						
Gemini-2.5-Flash	0.469	0.202	0.194	0.586	0.434	0.283
video-SALMONN-2	0.265	0.151	0.061	0.265	0.322	0.174
MiniCPM2.6-o	0.392	0.170	0.118	0.378	0.384	0.203
Qwen2.5-Omni	0.397	0.164	0.166	0.570	0.395	0.121

and 0.434 for open-ended QA on UGFV videos, compared to 0.422 and 0.386, respectively, on CCSF videos. These findings suggest that MLLMs face greater difficulty in comprehending humor in historically distant videos, *highlighting the sensitivity of humor understanding to the temporal and cultural context of videos.*

Table 6: Comparison between MLLMs and their base LLMs under the Text-Only setting.

Models	Open-ended QA		
	SentBERT	METEOR	BERTScore
Qwen2.5-VL	0.760	0.598	0.730
Qwen2.5-72B	0.692	0.624	0.690
Qwen2.5-Omni	0.719	0.540	0.687
Qwen2.5-7B	0.674	0.526	0.657

MLLMs vs. their base LLMs. MLLMs are usually derived from a pre-trained base LLM by adding a visual encoder or multimodal modules. For instance, Qwen2.5-VL extends Qwen2.5-72B, and Qwen2.5-Omni extends Qwen2.5-7B (see Table 6). In the Text-Only setup, Qwen2.5-Omni surpasses Qwen2.5-7B with a SentBERT score of 0.719 (vs. 0.674) and a BERTScore score of 0.687 (vs. 0.657) on open-ended QA task, suggesting that *multimodal training can confer advantages even when only textual descriptions are available, possibly because the model has learned richer contextual associations during training*. Please refer to Table 11 for more details on humor explanation and caption matching tasks.

5 Related Work

From LLMs to Video LLMs. Large language models have demonstrated outstanding capabilities in many domains, including natural language processing, coding, math, and reasoning, ushering in new breakthroughs for video understanding technology. Video LLMs integrate visual encoders with LLMs, leading to a unified model to reason

across video and language in the same language space (Wang et al., 2024b; Liu et al., 2024; Lin et al., 2023). Early video LLMs employ pre-trained image encoder and video encoder to encode only video frames (Zhang et al., 2023; Maaz et al., 2023; Li et al., 2023; Lin et al., 2023). Recent works augment video LLMs with an audio encoder to align visual, auditory, and textual modalities in the same language space. Moreover, the audio encoder is supposed to capture diverse environmental sound apart from human speech since sound has been shown to contain amounts of commonsense knowledge (Cheng et al., 2024; Xu et al., 2025).

Video LLMs. Video LLMs have shown remarkable performance in many traditional video processing tasks such as video captioning (Xu et al., 2016; Agrawal et al., 2019; Plummer et al., 2017), video question answering (Antol et al., 2015; Xiao et al., 2021; Yu et al., 2019; Fu et al., 2025), and grounding (Kazemzadeh et al., 2014; Wu et al., 2022). However, most existing benchmarks primarily target general video understanding tasks, such as MVBench (Li et al., 2024), Video-MME (Fu et al., 2025), PerceptionTest (Patraucean et al., 2023), MLVU (Zhou et al., 2025), and LVBench (Wang et al., 2024c), which mainly assess the recognition of basic visual cues across videos of varying lengths. Others are designed to evaluate specific video understanding capabilities, including temporal grounding (Gao et al., 2017; Lei et al., 2021; Hendricks et al., 2017; Wang et al., 2024d), video object detection (Shang et al., 2019, 2017), and video hallucination (Wang et al., 2024e; Leng et al., 2024). But there remains a pressing need for benchmarks that evaluate higher-level cognitive abilities, such as social intelligence, to better measure the gap between human and MLLMs' performance.

Our work narrows this gap. We expand the evaluation spectrum of video LLMs by introducing a challenging evaluation framework for humor understanding, formulating a humor generation task, and presenting the first comprehensive humor-centric evaluation and analysis.

Humor Video Understanding. Humor understanding is a popular research topic of artificial intelligence and its roots in cognitive science (Hampes, 2001, 2010). While early works focus on verbal humor like jokes and sarcasm, (Chłopicki, 2005; Petrović and Matthews, 2013; Joshi et al., 2017), the advent of LLMs has extended this scope to multimodal humor with image-language and video-language humor (Hessel et al., 2022; Alnajjar et al., 2022). However, a parallel evaluation of video LLMs is still limited.

While there have been several video-based humor datasets, the humor within them either is primarily dominated by spoken dialogue (Kumar et al., 2022; Hyun et al., 2024) or is restricted to those that can only be understood when both visual and linguistic cues are present (Ko et al., 2023), ignoring the fact that humans can appreciate humor solely from visuals. HumorQA (Xie et al., 2024) is most similar to ours, but it is restricted to surprising videos only and has an average duration of 7 seconds; consequently, it is limited in coverage and temporal complexity and is insufficient for humor-centric video understanding evaluation. Moreover, it overlooked environmental sound that generally enhances humor. In contrast, we focus on humor understanding and cover more diverse video humor, with an average duration twice that of HumorQA.

Sound has been found informative of commonsense (Zhao et al., 2022; Zellers et al., 2022), and it has been shown that integrating textual, acoustic, and visual features significantly improves humor detection accuracy (Chandrasekaran et al., 2016; Hasan et al., 2019). Since multimodal LLMs have recently been extended to natively support audio processing (aka. OmniLLMs), we propose and conduct a first evaluation of MLLMs on video humor understanding that involves environmental sound.

6 Conclusion

We have introduced v-HUB, a video humor understanding benchmark from vision and sound. v-HUB is designed to assess and diagnose the capability of MLLMs for video humor understanding. Each video is annotated with captions, descriptions, explanations, etc., supporting evaluation tasks such as caption matching, humor explanation, and open-ended QA. We evaluated a diverse range of MLLMs, spanning open-sourced and proprietary domains and covering specialized video LLMs and versatile OmniLLMs. Our findings reveal that current MLLMs heavily rely on linguistic cues for humor understanding, but are weak in deriving nuanced visual cues for understanding sophisticated video humor. Moreover, we empirically find that including environmental sound helps with humor understanding, highlighting the informativeness of sound and the promise of incorporating rich modalities for complex video reasoning tasks.

Limitations

We acknowledge two limitations of this work: (1) v-HUB contains humorous videos from diverse cultural backgrounds. Although culture-level categorization would be more fine-grained, assigning a single culture label to visual-centric humorous videos is often ambiguous and subjective, due to the prevalence of hybrid and globally shared visual humor. We therefore adopt region-level categorization, which is proposed by House et al. (2004) and Ronen and Shenkar (2013), as a more stable, weakly supervised proxy for cultural background, while acknowledging that a geographical region may encompass multiple cultures. (2) Second, all data annotations were created manually in English. However, the ability of LLMs to understand humor may vary across languages, as different languages interpret and express the same visual content in distinct ways. This points to a promising direction for future research.

Ethical Considerations

We firmly adhere to the ACL Code of Ethics in the performance of this work and the methods involved. We respect intellectual property and privacy rights and restrict the dataset to non-commercial academic use (see more details in Appendix A.3). Additionally, we carefully screened user-generated content to ensure the benchmark is safe for the research community and minimizes privacy risks for private individuals. Moreover, LLMs may produce offensive and incorrect statements, explicitly warn against the misuse of this dataset for generating malicious or mocking content, advocating instead for the development of safe and empathetic AI systems. This work is released with the intent for research purposes only.

Acknowledgements

Yanpeng Zhao acknowledges the support of the National Natural Science Foundation of China (12574467). Zilong Zheng is supported by the National Natural Science Foundation of China (62376031). We would like to thank Hengli Li for their helpful suggestions. We are also grateful to all annotators who contributed to the construction and verification of v-HUB. Their help and efforts were essential to ensuring the quality and reliability of the benchmark.

Use of LLMs. In this work, we used LLMs as assistive tools in the following stages:

- **Dataset Construction.** We initially employed GPT-4o (Hurst et al., 2024) to assist in generating candidate QA pairs and humor categories from video content. All outputs were subsequently reviewed and revised by human annotators to ensure correctness and quality.
- **Code Assistance.** LLMs were used to help generate parts of the evaluation code, which were then verified and refined by the authors.
- **Writing Support.** ChatGPT was used to write and polish some sentences in Section 4 for readability.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Khalid Alnajjar, Mika Hämäläinen, Jörg Tiedemann, Jorma Laaksonen, and Mikko Kurimo. 2022. When to laugh and how hard? a multimodal approach to detecting humor and its intensity. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6875–6886, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Moniek Buijzen and Patti M Valkenburg. 2004. Developing a typology of humor in audiovisual media. *Media psychology*, 6(2):147–167.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2016. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4603–4612.

Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024. Talk funny! a large-scale humor response dataset with chain-of-humor interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17826–17834.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476.

Władysław Chłopicki. 2005. The linguistic analysis of jokes: Graeme ritchie, routledge, london, 2004, 244 pp., hardback, £60. *Journal of Pragmatics*, 37(6):961–965.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society.

William P. Hampes. 2001. Relation between humor and empathic concern. *Psychological Reports*, 88(1):241–244. PMID: 11293036.

William P. Hampes. 2010. The relation between humor styles and empathy. *Europe's Journal of Psychology*, 6(3):34–45.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. 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. Association for Computational Linguistics.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5804–5813. IEEE Computer Society.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2022. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*.

Robert J House, Paul J Hanges, Mansour Javidan, Peter W Dorfman, and Vipin Gupta. 2004. *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2024. SMILE: Multimodal dataset for understanding laughter in video with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1149–1167, Mexico City, Mexico. Association for Computational Linguistics.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Dayoon Ko, Sangho Lee, and Gunhee Kim. 2023. Can language models laugh at youtube short-form videos? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 5956–5968, Dublin, Ireland. Association for Computational Linguistics.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *CoRR*, abs/2107.09609.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *CoRR*, abs/2410.12787.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *ArXiv*, abs/2305.06355.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *ArXiv*, abs/2311.10122.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2024. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, ArXiv:2306.05424.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, and 5 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Saša Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Sofia, Bulgaria. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vision*, 123(1):74–93.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Simcha Ronen and Oded Shenkar. 2013. Mapping world cultures: Cluster formation, sources and implications. *Journal of International Business Studies*, 44(9):867–897.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*, pages 279–287. ACM.

Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1300–1308. ACM.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-SALMONN 2: Captioning-Enhanced Audio-Visual Large Language Models. *arXiv preprint arXiv:2506.15220*.

Gemini 2.5 Team, Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and Others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, and 1 others. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. 2024a. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024c. Lvbench: An extreme long video understanding benchmark. *CoRR*, abs/2406.08035.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. 2024d. Hawkeye: Training video-text llms for grounding text in videos. *CoRR*, abs/2403.10228.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024e. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *CoRR*, abs/2406.16338.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2024. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16375–16387.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *Preprint*, ArXiv:2306.02858.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yanpeng Zhao, Jack Hessel, Youngjae Yu, Ximing Lu, Rowan Zellers, and Yejin Choi. 2022. Connecting the dots between audio and text without parallel data through visual knowledge transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4492–4507, Seattle, United States. Association for Computational Linguistics.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. MLVU: benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 13691–13701. Computer Vision Foundation / IEEE.

Overview of Appendix

This appendix provides supplementary materials that support the main paper. It includes (1) data curation and annotation procedures in Appendix A, (2) extended descriptions of task definitions and evaluation settings in Appendix B, (3) supplementary quantitative results in Appendix C, (4) qualitative case studies in Appendix D, (5) and the prompts used for evaluating MLLMs in Appendix E. These materials are intended to enhance the transparency, reproducibility, and completeness of the proposed benchmark and experimental findings.

A Crowdworking Details

A.1 Processing and Filtering

Harmful Content Detection. Before the annotation process began, we manually filtered out videos that contained potentially harmful content to ensure the video data's safety and quality (Figure 6 visualizes our annotation interface). Based on the criteria outlined by Thoppilan et al. (2022), we defined 6 categories of harmful contents, following aspects are checked for each video.

- *Discrimination.* Videos that display discrimination based on race, gender, sexual orientation, age, disability, appearance (e.g., obesity), or religion.
- *Animal Cruelty.* Videos that depict the abuse or mistreatment of animals.
- *Dangerous Activities.* Videos that include dangerous content such as drug use, criminal behavior, bullying, terrorism, rumor propagation, incitement, or misinformation.
- *Physical Violence.* Videos containing acts of physical violence against individuals, including fighting, severe injuries, bleeding, self-harm, or torture.
- *Obscenities.* Videos that contain explicit language, sexual behavior, or suggestive content.
- *Shocking Content.* Videos that include startling or fear-inducing elements such as gunshots, explosions, or jump scares.

In addition to harmful content detection, videos are also evaluated based on their quality:

- *Confusing:* Videos that are incomplete or otherwise difficult to understand.
- *Low Resolution:* Videos with a level of clarity that makes it challenging to discern the content.

Chaplin Video Segmentation. We selected 62 silent films by Charlie Chaplin and hired annotators to meticulously review each film, manually recording humorous moments to ensure each mime clip illustrates a whole mime through a single event or multi events. And we removed videos where both the reason for the humor and the action were repetitive (e.g. humor arising from a comical action due to inflexibility, such as failing to position a ladder properly) to ensure the quality and consistency of the videos and their annotations.

Speech Reliance Minimization. To ensure reliable identification of humorous content, we instructed two annotators to independently review each video and confirm the presence of clear humor. Each annotator was also instructed to review each video and label whether humor was primarily conveyed through visual cues and could be understood independently of speech. Only videos for which both annotators agreed were retained for the final dataset. We further employed Whisper (Radford et al., 2023), a performant speech-to-text model, to transcribe audio. Since Whisper transcribes filler sounds (e.g., "uh," "hmm") and other minimal utterances, we excluded any videos where the transcribed text exceeded 10 characters. Additionally, videos containing non-English speech were retained but muted, removing dependence on linguistic cues.

A.2 Annotation

Annotator Training. We provided appropriate annotation training for crowdworkers, offering detailed explanations of the annotation platform's usage and the annotation guidelines for different tasks. Additionally, we supplied an annotation manual (Figure 8) and corresponding instructional videos, which included specific descriptions and examples of the annotation requirements for crowdworkers to consult at any time during the annotation process.

Qualification. The recruited crowdworkers were mainly from China, all possessing at least an undergraduate education and with English background. Before formal annotation began, we conducted training sessions and a qualification review. During the qualification stage, crowdworkers were required to annotate 15 video samples. We manually reviewed their results and assigned scores based on the annotation guidelines. Ultimately, we selected eight qualified annotators. And we provided fair compensation to all crowdworkers, ensuring their hourly wages exceeded the local minimum wage.

For the annotation process, we adopted a three-person collaborative annotation scheme, ensuring that each data entry underwent three rounds of annotation. First, an annotator performed the initial annotation. Next, a second annotator reviewed and supplemented the annotation. Finally, a third annotator reviewed and further refined the previous two rounds of annotations. The annotators rotated through these three roles, and each annotation round was tracked to ensure that the three rounds

(a) **Visuals.** A man placed a battery on the conveyor belt, but it rolled against the belt's motion, forcing the cashier into an endless wait. For those who know the physics of a rolling cylinder on a moving conveyor, the scene feels even more clever.

(b) **Visuals**+Text. The video shows an animal rescue, with a cow dangling beneath a helicopter, appearing to swirl midair. The scene seems routine at first, but the added text 'milkshakes' cleverly parallels the moment, making it unexpectedly witty.

Figure 5: Examples of visual-centric humor understanding, where 'text' refers to visual text.

for each data entry were completed by different annotators. For humor rating and video captions, annotators were required to independently provide their own answers. For the remaining annotation tasks, when the second and third annotators reviewed and modified the previous annotations, they were required to submit a new annotation if they identified any issues. If a specific annotation issue was modified in all three rounds for a given video, we conducted a final review to assess the validity of the annotation results.

A.3 Copyright & License

We respect the copyright of each video. We have emailed Charlie Chaplin's copyright holders regarding copyright issues related to Chaplin clips, and v-HUB is only used for academic research. Commercial use in any form is prohibited. The copyright of all videos belongs to the video owners, and we will remove the videos upon their request. Without prior approval, you cannot distribute, publish, copy, disseminate, or modify v-HUB in whole or in part. You must strictly comply with the above restrictions.

B Additional Experimental Notes

B.1 Details of Generate Open-ended QA Pairs

We employed GPT-4o to generate QA pairs for each video, with the questions primarily covering temporal, descriptive, and causal aspects. The specific prompts used for QA generation are provided in Table 16. Subsequently, annotators manually reviewed and revised the QA pairs for each video to ensure their accuracy and quality.

Table 7: Evaluated Models.

Models	#Parameter	Proprietary	Input Modality		
			Text	Video	Video+Audio
Qwen2.5-VL	72B	✗	✓	✓	✗
Qwen2.5-Omni	7B	✗	✓	✓	✓
Intern3.5-VL	8B	✗	✓	✓	✗
MiniCPM2.6-o	8B	✗	✓	✓	✓
Video-SALMONN-2	7B	✗	✓	✓	✓
GPT-4o	-	✓	✓	✓	✗
Gemini-2.5-Flash	-	✓	✓	✓	✓

B.2 Details of Evaluation Methods

We used the 3.9.1 NLTK, 0.3.13 bert-score, and 5.1.1 sentence-transformers package to calculate the metrics.

AutoDQ. It evaluates the presence of humor-related events in the generated explanations (Wang et al., 2024a). It extracts key events from the model's output and compares them to ground truth (GT) annotations using entailment analysis. It provides three metrics: recall, precision, and F1 score, defined as:

- *Recall* measures the percentage of GT events entailed by the model-generated events.
- *Precision* measures the percentage of model-generated events that are entailed by GT events.
- *F1 Score* is the harmonic mean of precision and recall, serving as a balanced indicator that jointly reflects coverage and correctness.

The inclusion of AutoDQ allows us to evaluate factual correctness and event coverage, checking whether the explanations cover all the humorous points in the video content.

Human Evaluation. We randomly sampled 50 explanations generated by models for human evaluation. To ensure consistency in the evaluation criteria, we assigned one annotator to rate the humor explanations generated by models. The scores ranged from 0 to 100, and were subsequently normalized by a factor of 100, yielding final results within the range $[0, 1]$.

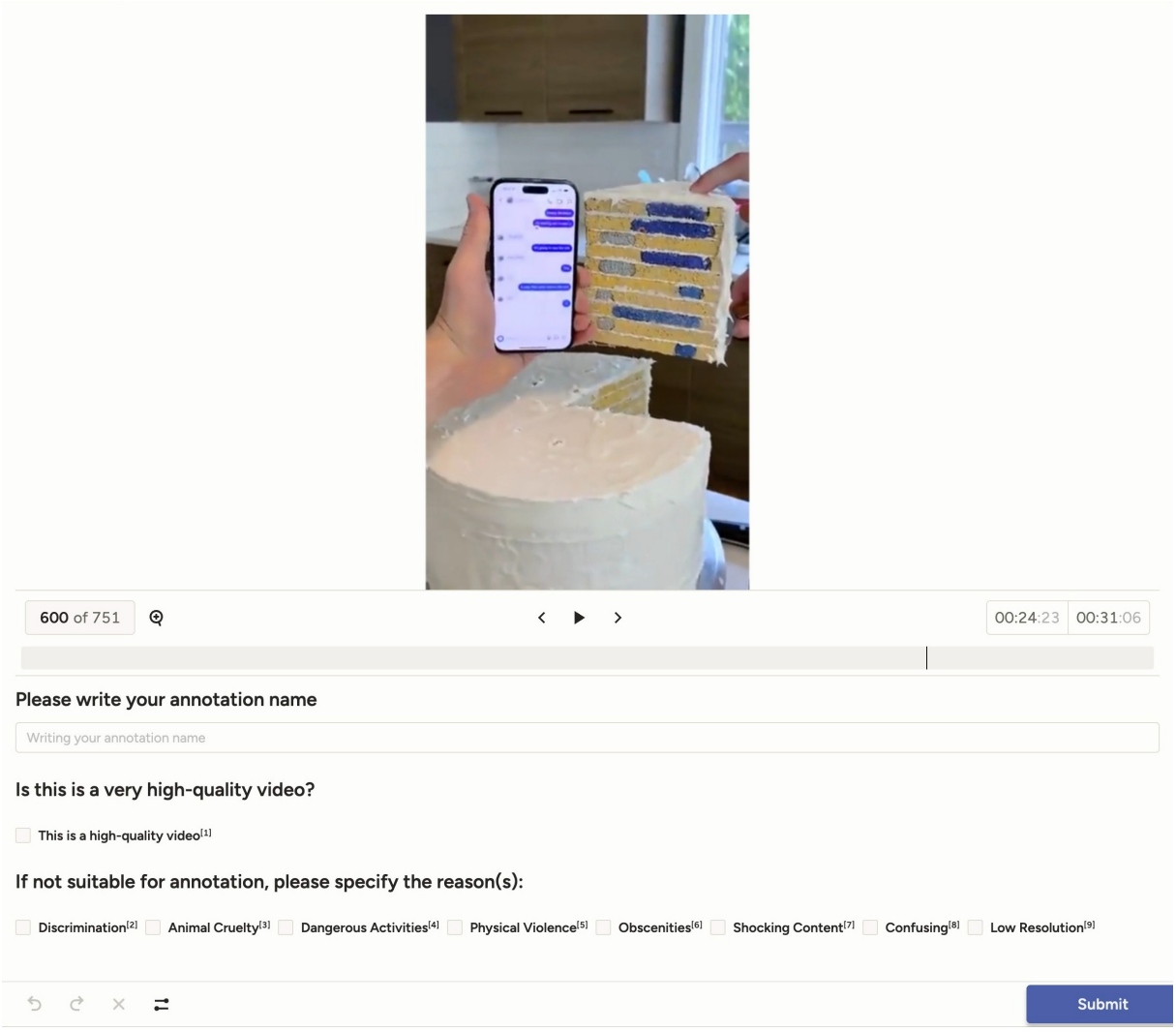


Figure 6: Interface for the Harmful Content Detection HIT.

118 of 274

00:04:21 00:11:09

For more details, see the Labeling Manual.

Please write your annotation name

Writing your annotation name

Q1. Please rate how humorous you find the video:

Very Humorous[1]

Somewhat Humorous[2]

Little bit Humorous[3]

Can't find any Humorous point[4]

undefined you can't find any humorous points, please skip to the next one.

Q2. Does the video contain any speech subtitles or visual text?

Speech Subtitles Exist[5]

Visual Text Exists[6]

Q3. Use your imagination to write a caption for the video that adds a new point of humor.

Write a humorous caption for the video

Q4. Pretend you're posting this video online. What caption will you write to make it more entertaining for viewers?

Write an entertaining caption for the video

Q5. Which aspects of the video contributed to understanding humor?

If a certain aspect appears in the video but does not help in understanding the video, you do not need to select it.

Human Visuals[7] Other Visuals (Objects or Scenery)[8] Speech (Spoken Words)[9] Visual Text[0] Speech Subtitles[q] Sound Effects or Music[w] Visual Effects[e]

Q6. Please describe the video in a direct way.

Detailing the main actions, people, and events without interpretation.

Enter a straightforward description of the video content

Q7. Please provide keywords / phrases representing background knowledge not shown in the video that an AI would need to understand the humor.

Include only knowledge that CANNOT be directly obtained by watching the video.

Enter keywords or phrases

Q8. Using your answers from Questions 5 to 7, please explain why the video is humorous.

Please include all the humorous points you can find, as thoroughly as possible.

Explain why the video is funny or interesting

Submit

Figure 7: Interface for HIT.

Annotation Manual

Please write your annotation name:

1. Please rate how humorous you find the video:

- ☐ **Very Humorous**
- ☐ **Somewhat Humorous**
- ☐ **Little bit Humorous**
- ☐ **Can't find any Humorous point**

For this question, please rate the humor level based on your first impression of the video.

2. Does the video contain any speech subtitles or visual text?

- ☐ **Speech Subtitles Exists**
- ☐ **Visual Text Exist**

For this question please select whether subtitles and text information can be seen in the video.

1. Details of options:
   - **Speech Subtitles**: Subtitles at the bottom of the video transcribing dialogues, narrations, and other speech, generally consistent with the spoken words.
   - **Visual Text**: Text visible in the video apart from subtitles, including added text in the video, text on objects, etc.
2. Use clear recognition by the human eye as the standard. If the content is difficult to see clearly, you do not need to select it.

Choose one Question between Question 3, 4 to Answer:

3. Use your imagination to write a caption for the video that adds a new point of humor.

This question ask you to add a caption to the video to increase its humor.

1. The video caption must be humorous only when combined with the video; the humor should not be understood by reading the text or watching the video alone.
2. Please ensure it is related to the video content.
3. The caption should not exceed one sentence.

4. Pretend you're posting this video online. What caption will you write to make it more entertaining for viewers?

This question ask you to write a caption for the video from the perspective of the video publisher. The caption should be connected to the video.

1. The caption should emphasize or enhance the humor of the video as much as possible. As the video publisher, you want to attract viewers.
2. Please ensure it is related to the video content.
3. The caption should not exceed one sentence.

5. Which aspects of the video contributed to understanding humor?

- ☐ **Visual - Human**
- ☐ **Visual - Others**
- ☐ **Visual Effects**
- ☐ **Visual Text**
- ☐ **Speech Subtitles**
- ☐ **Speech**
- ☐ **Sound Effects**

This question ask you to based on your understanding of the humor in the video, select the sources of information necessary for understanding the video and its humor.

1. If a certain aspect appears in the video but does not help in understanding the video, you do not need to select it.
2. The following are detailed explanations of the above categories:
   - **Visual - Human**: Information about human activities seen, including human expressions, actions, etc.
   - **Visual - Others**: Information seen other than human activities, including objects, backgrounds, creatures, etc.
   - **Visual Effects**: Post-production effects in the video, including special effects, filters, editing, etc.
   - **Visual Text**: Text seen in the video apart from subtitles, including added text in the video, text on objects, etc.
   - **Speech Subtitles**: Subtitles at the bottom of the video transcribing dialogues, narrations, and other speech, generally consistent with the spoken words.
   - **Sound Effects**: Audio information heard, including music, sound effects, meaningless shouts, exclamations, etc.
   - **Speech**: Spoken words heard, including dialogues, narrations, etc.

6. Please describe the video in a direct way. (Detailing the main actions, people, and events without interpretation)

Please describe what is happening in the video based only on what you see, including all the details necessary to understand the humor.

1. Please only describe the things that appear in the footage; do not make any inferential descriptions.
2. You may consider including:
   - Where does the video take place? Are there any changes in the scene?
   - Who appears in the video, and what are they doing?
   - What objects in the video need attention?
   - What are the expressions of the people in the video?

7. Please provide keywords / phrases representing background knowledge not shown in the video that an AI would need to understand the humor.

The question requires you to analyze the background knowledge necessary for understanding the video.

1. Please include only knowledge that cannot be directly obtained by watching the video. If there is none, you do not need to answer.
2. Please ensure it is directly related to understanding the humor.
3. Please be as specific as possible. For example: "5G" is better than "Networks", "John F. Kennedy" is better than "US President".

8. Using your answers from Questions 5 to 7, please explain why the video is humorous.

The question requires you to explain the humor in the video.

1. Please answer based on your analysis of the video content in questions 5, 6, and 7. For example, if in question 5 you selected 'sound effect', explain why the sound effect makes people feel humor.
2. Please include all the humorous elements you can find, as thoroughly as possible.

Figure 8: Interface for Annotation Manual for data annotation.

Table 8: The impact of requiring background knowledge support on video humor understanding in the Video-Only setting.

Models	Explanation			Matching	Open-ended QA	
	SentBERT	METEOR	AutoDQ	Accuracy	SentBERT	METEOR
Background-Dependent videos						
Gemini-2.5-Flash	0.500	0.211	0.198	0.628	0.433	0.266
video-SALMONN-2	0.271	0.153	0.028	0.257	0.307	0.160
MiniCPM2.6-o	0.402	0.166	0.093	0.374	0.328	0.120
Qwen2.5-Omni	0.401	0.158	0.137	0.559	0.382	0.105
Full dataset						
Gemini-2.5-Flash	0.459	0.199	0.175	0.580	0.424	0.270
video-SALMONN-2	0.269	0.150	0.052	0.243	0.317	0.169
MiniCPM2.6-o	0.381	0.165	0.112	0.362	0.369	0.186
Qwen2.5-Omni	0.384	0.159	0.144	0.553	0.382	0.121

B.3 Baseline Models

To evaluate multimodal large language models' ability to understand video humor, we selected state-of-the-art models representing three distinct input modalities, as summarized in Table 7. Specifically, we include multimodal LLMs that process raw visual frames and text, and omni LLMs that integrate both text, video and audio signals. This set covers both public models (e.g., Qwen2.5-VL, Intern3.5-VL) and proprietary models (e.g., Gemini-2.5-Flash, GPT-4o), offering a broad perspective on current approaches. Each model is evaluated under all input conditions it can handle (see Section 4.1): for instance, omni-modal models can participate in the Text-Only, Video-Only, and Video+Audio groups, whereas multimodal models are tested exclusively with textual input and raw visual frames. This setup allows us to isolate how each model category—multimodal and omni-modal—contributes to humor understanding across diverse input modalities.

C Additional Experimental Results

Comparable performance on videos with and without background knowledge requirements. The results in Table 8 show that model performance on Background-Dependent videos is largely comparable to that on the full dataset in the Video-Only setting. For example, video-SALMONN-2 attains an average SentBERT of 0.271 on the humor explanation task for background-dependent videos, which is statistically similar to its SentBERT of 0.269 on the full dataset. This suggests that the language-model component of MLLMs already encodes most of the cultural background knowledge necessary for humor comprehension, meaning that *the absence of explicit background knowledge in the input does not significantly degrade their performance*. MLLMs show a comparable performance in understanding videos that require background knowledge compared to those that do not, potentially because video humor rarely relies on specific background knowledge, making it universally understandable.

Table 9: Human preference comparison of humor explanations across four model categories.

Models	Proprietary	Type	Setting		
			Text-Only	Video-Only	Video+Audio
Qwen2.5-VL	✗	MLLM	0.687	0.423	–
Qwen2.5-Omni	✗	OmniLLM	0.574	0.430	0.381
GPT-4o	✓	MLLM	0.654	0.576	–
Gemini-2.5-Flash	✓	OmniLLM	0.651	0.546	0.566

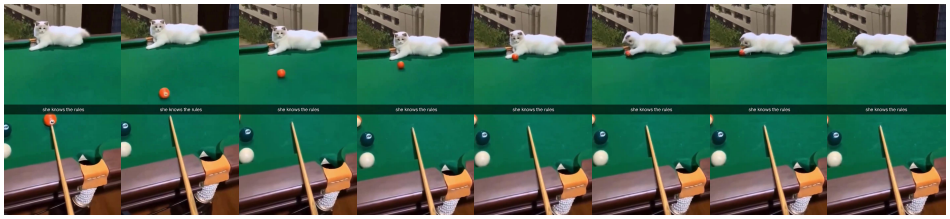
Proprietary MLLMs show stronger resilience to multimodal inputs compared to public MLLMs. The results in Table 9 indicate that current MLLMs rely heavily on linguistic cues to generate reasonable explanations, and struggle to effectively extract semantic information from raw visual or auditory signals. For example, Qwen2.5-VL attains a preference score of 0.687 under Text-Only, significantly outperforming its Video-Only score of 0.423. Furthermore, although closed-source models demonstrate greater robustness under multimodal inputs, they still struggle to align visual and audio cues to enhance humor comprehension. For instance, Gemini-2.5-Flash achieves 0.546 (Video-Only) and 0.566 (Video+Audio).

D Case Study

We present our case studies in Figure 9.

E Prompts

We list our prompt in Tables 16 to 27



Task 1: Caption Matching

You will see five captions, pick the only one that best corresponds to the humor in the video. The correct caption may relate to the humor either directly or indirectly.

A. Is the moving-to-moving parabolic projection so accurate?
B. Very quick response
C. Probably the last time I played it.
D. No moral ethics!
E. The problem is that you don't use a white ball. ✓

Model Answer:

Text-Only: E Video-Only: E Video-Audio: E

Task 2: Open-ended QA

Q: What did the kitten do with the ball?
A: Caught it with its claws and put it into the hole.

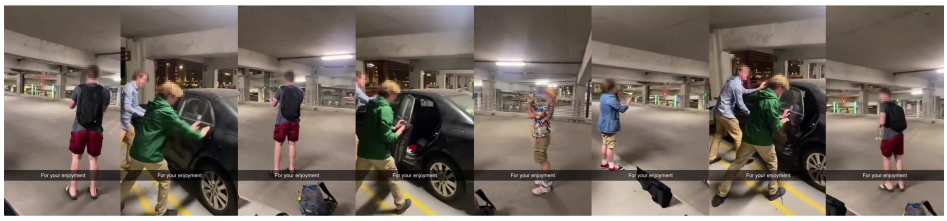
Model Answer:

Text-Only: The kitten **caught the ball** with its claws and **put it into the hole**.

Video-Only: The kitten **didn't interact with the ball**; it was just present in the scene.

Video-Audio: **pushed it**.

(a) case study 1



Task 1: Caption Matching

You will see five captions, pick the only one that best corresponds to the humor in the video. The correct caption may relate to the humor either directly or indirectly.

A. Not a human at all
B. The big dumbbell of the skinny man
C. Where else are you hiding?
D. The ball goes in!
E. Watch this epic parking garage symphony featuring trombones and car doors—pure chaotic genius! ✓

Model Answer:

Text-Only: E Video-Only: B Video-Audio: E

Task 2: Open-ended QA

Q: What are the two people on the side of the car doing?
A: Holding and slamming the car doors.

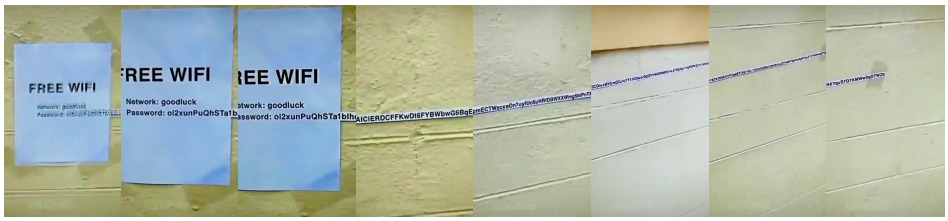
Model Answer:

Text-Only: **Holding the car doors and slamming them** hard to make a sound in coordination with the horn, and then blowing the horn to join in the music.

Video-Only: **Talking**

Video-Audio: The two people on the side of the car are **pushing it**.

(b) case study 2



Task 1: Caption Matching

You will see five captions, pick the only one that best corresponds to the humor in the video. The correct caption may relate to the humor either directly or indirectly.

A. Really free ✓
B. Car window eating incident
C. Balenciaga is like this
D. Kick away
E. Putting this together is fitness

Model Answer:

Text-Only: A Video-Only: D Video-Audio: A

Task 2: Open-ended QA

Q: Why do few people use the free wifi?
A: The password is too long.

Model Answer:

Text-Only: The password is **too long**.

Video-Only: Wi-Fi is **not working**

Video-Audio: The password is very long and **hard to remember**.

(c) case study 3

Figure 9: Three case studies illustrating our findings.

Table 10: The impact of visual text on video humor understanding in the Video+Audio setting.

Models	Sound contributing to humor							Sound not contributing to humor						
	Explanation						Open-ended QA	Explanation						Open-ended QA
	SentBERT	METEOR	BERTScore	Precision	Recall	F1 Score	BERTScore	SentBERT	METEOR	BERTScore	Precision	Recall	F1 Score	BERTScore
w/ visual text														
Gemini-2.5-Flash	0.458	0.199	0.541	0.157	0.210	0.174	0.552	0.472	0.206	0.547	0.185	0.174	0.177	0.550
video-SALMONN-2	0.258	0.164	0.481	0.169	0.042	0.067	0.532	0.261	0.160	0.509	0.077	0.045	0.057	0.540
MiniCPM2.6-o	0.416	0.170	0.516	0.105	0.111	0.108	0.527	0.434	0.182	0.522	0.143	0.137	0.151	0.536
Qwen2.5-Omni	0.441	0.170	0.522	0.148	0.133	0.140	0.521	0.430	0.176	0.522	0.159	0.133	0.145	0.538
w/o visual text														
Gemini-2.5-Flash	0.482	0.202	0.550	0.165	0.195	0.179	0.539	0.448	0.195	0.543	0.034	0.051	0.041	0.552
video-SALMONN-2	0.276	0.185	0.496	0.124	0.045	0.064	0.529	0.285	0.170	0.504	0.098	0.034	0.050	0.541
MiniCPM2.6-o	0.407	0.182	0.525	0.130	0.097	0.103	0.515	0.364	0.167	0.510	0.060	0.048	0.046	0.514
Qwen2.5-Omni	0.425	0.181	0.538	0.116	0.128	0.122	0.509	0.413	0.173	0.522	0.062	0.045	0.050	0.529

Table 11: Comparison between MLLMs and their base LLMs under the Text-Only setting.

Models	Explanation						Matching
	SentBERT	METEOR	BERTScore	Precision	Recall	F1 Score	Accuracy
Qwen2.5-VL	0.543	0.250	0.573	0.323	0.364	0.342	0.719
Qwen2.5-72B	0.536	0.245	0.576	0.318	0.374	0.344	0.661
Qwen2.5-Omni	0.536	0.233	0.565	0.303	0.331	0.316	0.644
Qwen2.5-7B	0.560	0.240	0.567	0.293	0.369	0.324	0.542

Table 12: Model performance on Humor Explanation.

Models	Explanation				Open-ended QA
	BERTScore	Precision	Recall	F1 Score	BERTScore
Text-Only					
Gemini-2.5-Flash	0.575	0.307	0.385	0.342	0.712
video-SALMONN-2	0.586	0.290	0.350	0.317	0.639
MiniCPM2.6-o	0.558	0.307	0.345	0.325	0.536
Qwen2.5-Omni	0.565	0.303	0.331	0.316	0.687
Qwen2.5-VL	0.573	0.323	0.364	0.342	0.730
Intern3.5-VL	0.576	0.320	0.382	0.348	0.685
GPT-4o	0.574	0.339	0.417	0.374	0.699
Video-Only					
Gemini-2.5-Flash	0.546	0.154	0.206	0.176	0.549
video-SALMONN-2	0.497	0.087	0.042	0.052	0.525
MiniCPM2.6-o	0.516	0.116	0.109	0.112	0.469
Qwen2.5-Omni	0.518	0.169	0.126	0.144	0.497
Qwen2.5-VL	0.542	0.150	0.152	0.150	0.546
Intern3.5-VL	0.537	0.126	0.116	0.125	0.540
GPT-4o	0.536	0.214	0.205	0.206	0.544
Video+Sound					
Gemini-2.5-Flash	0.546	0.153	0.200	0.173	0.549
video-SALMONN-2	0.499	0.126	0.047	0.066	0.537
MiniCPM2.6-o	0.519	0.122	0.062	0.110	0.514
Qwen2.5-Omni	0.525	0.137	0.116	0.120	0.529

Table 13: The impact of requiring background knowledge support on video humor understanding in the Video-Only setting.

Models	Explanation				Open-ended QA
	BERTScore	Precision	Recall	F1 Score	BERTScore
Background-Dependent videos					
Gemini-2.5-Flash	0.566	0.193	0.203	0.198	0.549
video-SALMONN-2	0.512	0.051	0.019	0.028	0.488
MiniCPM2.6-o	0.531	0.109	0.079	0.093	0.446
Qwen2.5-Omni	0.530	0.184	0.109	0.137	0.519
Full dataset					
Gemini-2.5-Flash	0.546	0.154	0.206	0.176	0.549
video-SALMONN-2	0.497	0.087	0.042	0.052	0.525
MiniCPM2.6-o	0.516	0.116	0.109	0.112	0.469
Qwen2.5-Omni	0.518	0.169	0.126	0.144	0.497

Table 14: The impact of background knowledge on video humor understanding in the Video+Audio setting.

Models	Explanation				Open-ended QA
	BERTScore	Precision	Recall	F1 Score	BERTScore
w/ Background Knowledge					
video-SALMONN-2	0.562	0.117	0.107	0.114	0.536
MiniCPM2.6-o	0.555	0.192	0.194	0.193	0.520
Qwen2.5-Omni	0.557	0.195	0.160	0.176	0.555
w/o Background Knowledge					
video-SALMONN-2	0.514	0.084	0.014	0.025	0.528
MiniCPM2.6-o	0.538	0.132	0.103	0.115	0.509
Qwen2.5-Omni	0.544	0.142	0.114	0.127	0.525

Table 15: The impact of video era on video humor understanding in the Video-Only setting.

Models	Explanation				Open-ended QA
	BERTScore	Precision	Recall	F1 Score	BERTScore
Last-Century *Charlie Chaplin's Silent Films*					
Gemini-2.5-Flash	0.541	0.118	0.145	0.130	0.545
video-SALMONN-2	0.509	0.035	0.007	0.012	0.513
MiniCPM2.6-o	0.508	0.116	0.083	0.097	0.470
Qwen2.5-Omni	0.510	0.153	0.070	0.096	0.493
Contemporary *User-Generated Funny Video*					
Gemini-2.5-Flash	0.547	0.173	0.222	0.194	0.550
video-SALMONN-2	0.511	0.103	0.052	0.061	0.530
MiniCPM2.6-o	0.518	0.113	0.124	0.118	0.470
Qwen2.5-Omni	0.520	0.173	0.130	0.166	0.498

Table 16: Prompt for generate QA pairs.

```
These are frames from a video.
And you'll be given a description of a video and an explanation of why it's humorous to watch.
Based on given information, generate a Video Reasoning QA pair, try to make answer only as phrases. Let's think step by step. \n
Additionally, classify this question into one of the following categories using the concise definitions provided: \n
Descriptive question: Involves factual details such as location or count \n
Temporal question: Involves time-related aspects (e.g., previous, after) \n
Causal question: Involves reasons or explanations (e.g., why, how) \n\n
Example 1: \n
Description: \n
Two hands are stretched out, one hand holding KFC chicken nuggets and the other hand holding seeds. In the distance, a chicken runs over, but the chicken prefers to eat the KFC chicken. \n
Explanation: The chicken surprisingly likes to eat KFC chicken, which is unexpected and a bit funny. The man realizes something is wrong and tries to push the chicken pieces away with his hand, which adds to the humor with a sense of panic. \n\n
Question: What does the man holding in his hand? \n
Answer: KFC chicken nuggets and seeds. \n
Type: Descriptive \n\n
Example 2: \n
Description: A man poured red liquid into the water, and a group of fish came to snatch the food. Another man poured beer into the water, and a group of men came to snatch the food like fish. \n
Explanation: The portrait of people snatching food like fish humorously reflects the attraction of beer to men, and the connection between them is very funny. \n\n
Question: After the man poured beer into the water, what happened? \n
Answer: A group of men came. \n
Type: Temporal \n\n
Example 3: \n
Description: A woman was lying on the handrail of an escalator while moving down. A man saw her, and lying on the handrail on the other side, and as a result, there was no barrier on that side, and he fell directly down the escalator. \n
Explanation: The man tried to show off by imitating others, but ended up falling hard, which made people find it funny. \n\n
Question: Why does the man fall off on the other side of the handrail? \n
Answer: There was no barrier. \n
Type: Causal \n\n
Output format: \n
Question: <question> \n
Answer: <answer> \n
Type: <type> \n\n
Video Description: {video_description} \n
Humor Explanation: {humor_explanation} \n
```

Table 17: Prompt for video QA.

```
System: You are a helpful AI assistant. You can analyze videos and answer questions about their content. Respond with short and concise answers. Avoid using unpronouncable punctuation or emojis.

User: These are frames from a video. Based on these frames, answer the following question: {question} \n\n

Output format: \n
Answer: <answer> \n\n
```

Table 18: Prompt for video explanation.

```
System: You are a helpful AI assistant specialized in video understanding and humor analysis. You can explain jokes clearly and naturally based on video content and video description. Please respond with short and concise answers. Avoid using unpronouncable punctuation or emojis.

User: These are frames from a video. Your job is to explain why the video is humorous in 2-3 sentences as if you were explaning to a friend who doesn't get the joke yet. Respond with a 2-3 sentence explanation of the joke and how it relates to the video. \n\n

Output format: \n
Explanation: <answer> \n\n
```

Table 19: Prompt for video caption matching.

```
System: You are a helpful AI assistant. You can analyze videos and answer questions about their content. Please only output in the specified format. No extra text.

User: Along with the frames from the video. And {question} \n
Please respond with response with the option letter only. \n\n

Output format: \n
```

Table 20: Prompt for video with description QA.

```
System: You are a helpful AI assistant. You can analyze videos and answer questions about their content. Respond with short and concise answers. Avoid using unpronouncable punctuation or emojis.

User: You'll be given a description of the video. Based on this information, answer the following question: {question} \n\n

Output format: \n
Answer: <answer> \n\n

Video Description: {video_description}
```

Table 21: Prompt for video with description explanation.

```
System: You are a helpful AI assistant specialized in video understanding and humor analysis. You can explain jokes clearly and naturally based on video content and video description. Please respond with short and concise answers. Avoid using unpronouncable punctuation or emojis.

User: You will also be given a description of the video. Your job is to explain why the video is humorous in 2-3 sentences as if you were explaning to a friend who doesn't get the joke yet. Respond with a 2-3 sentence explanation of the joke and how it relates to the video. \n\n

Output format: \n
Explanation: <answer> \n\n

Video Description: {video_description}
```

Table 22: Prompt for writing captions of videos.

```
And I will provide a description of the video and a list of descriptive captions that break down what happens in it.
Your task is to write a caption in one sentences from the video creator's perspective — something you would write to attract viwers.
Requirements:
Please ensure it is related to the video content.
- Write as if you're sharing it with an audience (e.g., use 'this' or 'me' naturally).
Output format:
Caption: <caption>
Video description: {video_description}
Descriptive captions: {descriptive_captions}
```

Table 23: Prompt for video with description caption matching.

```
System: You are a helpful AI assistant. You can analyze videos and answer questions about their content. Please only output in the specified format. No extra text.

User: You'll be given a description of the video. And {question}\n Please respond with response with the option letter only.\n\n

Output format:\n
Answer: <answer>\n\n
Video Description: {video_description}
```

Table 24: Prompt for video with sound QA.

```
System: You are a helpful AI assistant. You can analyze videos and answer questions about their content. Respond with short and concise answers. Avoid using unpronouncable punctuation or emojis.

User: Here's a humorous video. Based on the its visual and audio information, answer the following question: {question} \n\n

Output format: \n
Answer: <answer> \n\n
```

Table 25: Video with sound explanation.

```
System: You are a helpful AI assistant specialized in video understanding and humor analysis. You can explain jokes clearly and naturally based on video content and video description. Please respond with short and concise answers. Avoid using unpronouncable punctuation or emojis.

User: Here's a humorous video. Your job is to explain why the video is humorous in 2-3 sentences as if you were explaning to a friend who doesn't get the joke yet. Respond with a 2-3 sentence explanation of the joke and how it relates to the video. \n\n

Output format: \n
Explanation: <answer> \n\n
```

Table 26: Video with sound caption matching

```
System: You are a helpful AI assistant. You can analyze videos and answer questions about their content. Please only output in the specified format. No extra text.

User: Along with visual and audio information in the video. And {question} \n
Please respond with response with the option letter only. \n\n

Output format: \n
Answer: <answer> \n\n
```

Table 27: Prompt for video with background knowledge caption matching

```
System: You are a helpful AI assistant. You can analyze videos and answer questions about their content. Respond with short and concise answers. Avoid using unpronounceable punctuation or emojis.

User: Here's a humorous video. You will be given background knowledge of the video. Based on its visual and audio information and the background knowledge, answer the following question: {question} \n
Please respond with the option letter only. \n\n

Output format: \n
Answer: <answer> \n\n
Background Knowledge: {background_knowledge}
```