😃 SMILE: Multimodal Dataset for Understanding Laughter in Video with Language Models

Lee Hyun[1,2*†] **Kim Sung-Bin**[1*] **Seungju Han**[3] **Youngjae Yu**[4] **Tae-Hyun Oh**[1,5,6]

[1]Dept. of Electrical Engineering and [5]Grad. School of Artificial Intelligence, POSTECH
[2]Samsung Advanced Institute of Technology
[3]Seoul National University [4]Yonsei University
[6]Institute for Convergence Research and Education in Advanced Technology, Yonsei University
{hyunlee, sungbin, taehyun.oh}@postech.ac.kr

Abstract

Despite the recent advances of the artificial intelligence, building social intelligence remains a challenge. Among social signals, laughter is one of the distinctive expressions that occurs during social interactions between humans. In this work, we tackle a new challenge for machines to understand the rationale behind laughter in video, *Video Laugh Reasoning*. We introduce this new task to explain why people laugh in a particular video and a dataset for this task. Our proposed dataset, 😃 SMILE, comprises video clips and language descriptions of why people laugh. We propose a baseline by leveraging the reasoning capacity of large language models (LLMs) with textual video representation. Experiments show that our baseline can generate plausible explanations for laughter. We further investigate the scalability of our baseline by probing other video understanding tasks and in-the-wild videos. We release our dataset, code, and model checkpoints on https://github.com/postech-ami/SMILE-Dataset.

1 Introduction

"Laughter is the shortest distance between two people."
—Victor Borge

We, human beings, are immersed in laughter. Laughter is a distinctive non-verbal social signal, associated with bonding, agreement, affection, and emotional regulation (Scott et al., 2014). It is often purposedly elicited to establish intimacy (Stauffer, 1999), grab attention (Wanzer et al., 2010), or build faith (Vartabedian and Vartabedian, 1993); *i.e.*, serving as a powerful medium to express a wide range of social and emotional implications beyond the capacity of mere words. Thus, understanding laughter is a crucial problem with huge potential in artificial social intelligence (Bainbridge et al., 1994; Williams et al., 2022; Dautenhahn, 2007) to build empathetic machines with human-machine interaction (Lee et al., 2017; Nijholt et al., 2017; Inoue et al., 2022). However, understanding and modeling laughter reactions is challenging. Even a simple joke is associated with language skills, context knowledge, theory-of-mind, abstract thinking, and social perception, and complex entanglement of these makes laughter reaction arguably the most complex cognitive attribute humankind may have (McDonald, 2013).

In this work, we take the first stepping stone to tackle the challenge of understanding laughter by introducing a task, *Video Laugh Reasoning* that aims to interpret the reasons behind laughter in a video. For this task, we curate a new dataset, 😃 SMILE, consisting of video clips and corresponding text annotations explaining reasons for laughter. We probe through the question "Why do people laugh?" and reason through the answer in a language form; thus, we define the task as a free-form text generation task in which the model generates an explanation for the laughter with a given video clip (See Figure 1).

While reasoning laughter by answering the question is an effective way of probing the level of understanding, laughter itself has an inherently complex nature which can be influenced by diverse factors (Apte, 1985; Provine, 2001; Martin et al., 2003; Martin and Ford, 2018), *e.g.*, the subjectivity (Warren et al., 2021), context knowledge (Nijholt et al., 2017), and multimodality (Hasan et al., 2019). To build a clearer resource for understanding laughter and its social norm behind it, we design the dataset to focus on *audience laughter*, a cohesive form from social influence in distinct contexts (Greatbatch and Clark, 2003), and thereby alleviating the subjectivity associated with individual laughter. Also, for our task, we propose a baseline based on large language models (LLMs) with *multimodal*

*equally contributed
†work done at POSTECH

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1149–1167

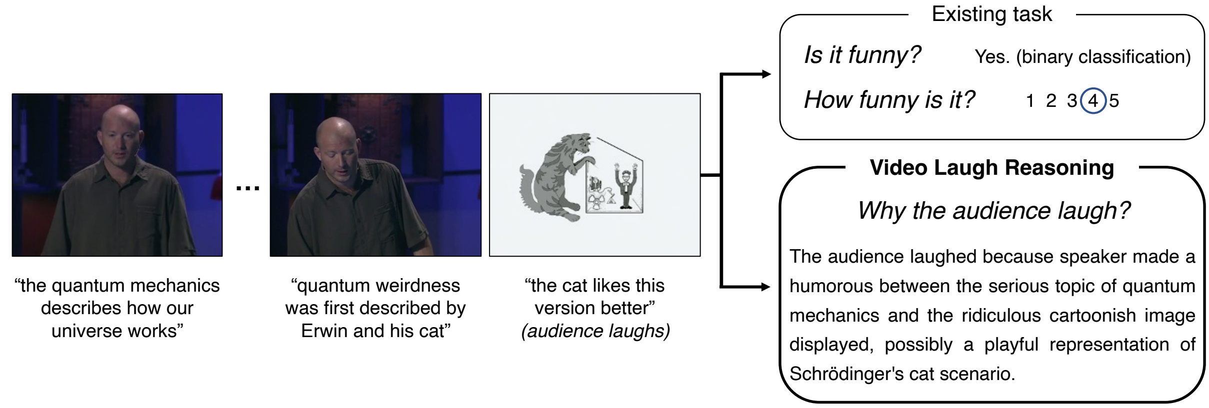


Figure 1: Why do people laugh? We present *Video Laugh Reasoning*, a new task to interpret the reasons behind laughter in a video.

textual representation by converting multimodal attributes and features on video into a textual format.

Our experimental results show that the proposed baseline, incorporating LLM's reasoning capability with multimodal textual representation, can generate plausible explanations of the reason for laughter. Our data analysis and ablation study reveals that multimodal information plays a role in understanding laughter. We further explore the scalability of utilizing LLM with textual representation by applying it to other video understanding tasks and in-the-wild videos.

Our major contributions are threefold: 1) proposing *Video Laugh Reasoning*, a new task for understanding the reason behind laughter in a video, 2) building 😀 SMILE, a new dataset that comprises video and explanation for laughter reason, and 3) presenting a baseline using LLM with multimodal textual representation for laugh reasoning task and its scalability.

2 Related Work

Understanding laughter. Laughter plays a key role in social interactions, such as bonding, agreement, affection, and emotional regulation (Scott et al., 2014). Given its importance in social interactions, seminar works tackle to detect laugh-inducing moments, specifically focusing on humor or sarcasm. Several methods (Annamoradnejad and Zoghi, 2020; Weller and Seppi, 2020) rely primarily on transcripts for humor detection. As laughter occurs with multimodal information, such as variations in tone or facial cues, there are attempts to incorporate audio and text cues from videos (Bertero and Fung, 2016; Alnajjar et al., 2022), or even include visual cues (Castro et al., 2019; Hasan et al., 2019; Ray et al., 2022) to pinpoint the occurrences of humor. Yet they focus on detecting whether a certain situation induces laughter or predicting the intensity of laughter, without providing explanations for the underlying reasons behind the laughter (See Figure 1). Moreover, despite the availability of datasets for understanding the types and characteristics of laughing moments (Urbain et al., 2010; McKeown et al., 2012; Dupont et al., 2016), no dedicated dataset is available for comprehending the context surrounding laughter. Few works (Chowdhery et al., 2022; Hessel et al., 2023; Ko et al., 2023) have attempted to reason about laughter or jokes. However, their scope differs from ours, as they focus on providing instant textual descriptions of humor or cartoon images accompanied by text. To the best of our knowledge, we are the first to introduce the task of understanding the reason for laughter within videos, accompanied by our comprehensive dataset.

Multimodal reasoning. Multimodal reasoning is a complex task aiming to equip machines with the capability to parse, analyze, and logically reason about the given multimodal context. A widely explored reasoning task is a question answering (QA) on images (Antol et al., 2015; Gao et al., 2015; Zhu et al., 2016) or video (Lei et al., 2018; Tapaswi et al., 2016), which requires understanding the question, referencing the appropriate context, and selecting the correct answer. Similarly, commonsense reasoning (Vedantam et al., 2015; Yatskar et al., 2016; Wu et al., 2016) is another type of reasoning, demanding a more profound level of understanding and the ability to infer unstated information. Our task includes commonsense reasoning in that laughter is often elicited by exploiting external contexts, rather than merely understanding underlying phenomena.

Several methods (Zellers et al., 2019; Vicol et al., 2018; Zadeh et al., 2019) have attempted to learn and reason about the social interactions in the video. For instance, Visual Commonsense Reasoning (VCR) (Zellers et al., 2019) unifies reasoning about diverse commonsense phenomena, while Social IQ (Zadeh et al., 2019) aims to teach social intelligence by providing a broad range of social and behavioral situations to a machine. However, these approaches give less attention to a deeper understanding of laughter itself—a complex nonverbal signal integral to social interactions. Unlike the prior arts, we specifically focus on the task of reasoning human laughter. We posit this as a significant stride towards understanding important social signals frequently encountered in daily life, thus contributing a new perspective to multimodal reasoning and understanding tasks.

Models for multimodal reasoning. To tackle multimodal reasoning, one approach is to design pretraining methods (Lu et al., 2019; Li et al., 2019) that learn the joint vision and language representations. More recently, the combination of large-scale vision and language models (VLM) has demonstrated remarkable performance in multimodal reasonings (Li et al., 2023; Lu et al., 2022; Zhang et al., 2023; Wang et al., 2022a; Han et al., 2023).

An alternative approach for multimodal reasoning utilizes text as a unified representation and large language models (LLM) with minimal or without training. For instance, Socratic Model (Zeng et al., 2022) employs language to combine complementary knowledge from various pre-trained models for tackling a wide range of tasks. Similarly, Wang et al. (2022c) converts the visual attributes into the text representation to prompt a frozen LLM for diverse video-language tasks. In this work, we conduct extensive experiments on our proposed laugh reasoning task and show the effectiveness of using text as an intermediate representation.

3 Task Definition and Dataset

In this section, we introduce our *Video Laugh Reasoning* task and our dataset for it.

3.1 Task Definition and Baseline

We present *Video Laugh Reasoning*, a task that challenges the model to understand reasons for laughter in a given video. We pose our task as a generation problem, enabling the model to explain why a particular situation incited laughter in the video. We define this task as, $\hat{y} = f(v)$, where $\hat{y}$, f, and v stand for the generated explanation about laughter reason, the model, and the given video clip.

For this task, we propose a baseline that utilizes the reasoning capacity of LLM. To ensure compatibility of input v with the language model, we convert videos into *multimodal textual representation* that preserve multimodal information from video, such as visual, acoustic, and semantic cues. We compose visual cues with facial expressions[1] and scene descriptions[2] to perceive human-specific and scene-wide contextual information. For acoustic cues, we extract the mean and the variance of pitch, intensity, jitter, and shimmer from speech to capture. We simply use transcripts of the speech from the videos for semantic cues (See Figure 2).

Using textual representation as input and LLM as model f, we can rewrite the task formula as, $\hat{y} = f(\mathcal{P}, \{t_1, t_2, ..., t_k\})$, where $\mathcal{P}$ stand for the prompt that describes input representation and instructing the laugh reasoning task to language models and t is multimodal textual representation converted from the given video clip v. See Appendix A for details about how to convert video into textual representation.

3.2 Dataset

Data collection. We present 😀 SMILE, a curated dataset encompassing 887 video clips, each paired with a language description about the reason for laughter for the corresponding video clip. This pairing facilitates supervised training for the laugh reasoning task. The dataset focuses on audience laughter among many types of laughter since audience laughter usually has a clearer signal than other laughter and represents a general and cohesive form of laughter. To encompass a wider range of videos that contain situations where audiences laugh, we construct our dataset using two different sources: *TED talks and sitcoms*.[3]

We curate video clips that span between 10 and 90 seconds for *TED talks* and 7 and 60 seconds for *sitcoms*. If a video is too short, it might fail to provide sufficient context for laughter. In contrast, if a video is too long, it may dilute specific laughter-inducing contexts with unrelated information. The

[1]We use facial action units (Ekman and Friesen, 1978).

[2]We use video captioning model (Wang et al., 2022b).

[3]We source the video clips from MUStARD (Castro et al., 2019) and UR-Funny dataset (Hasan et al., 2019).

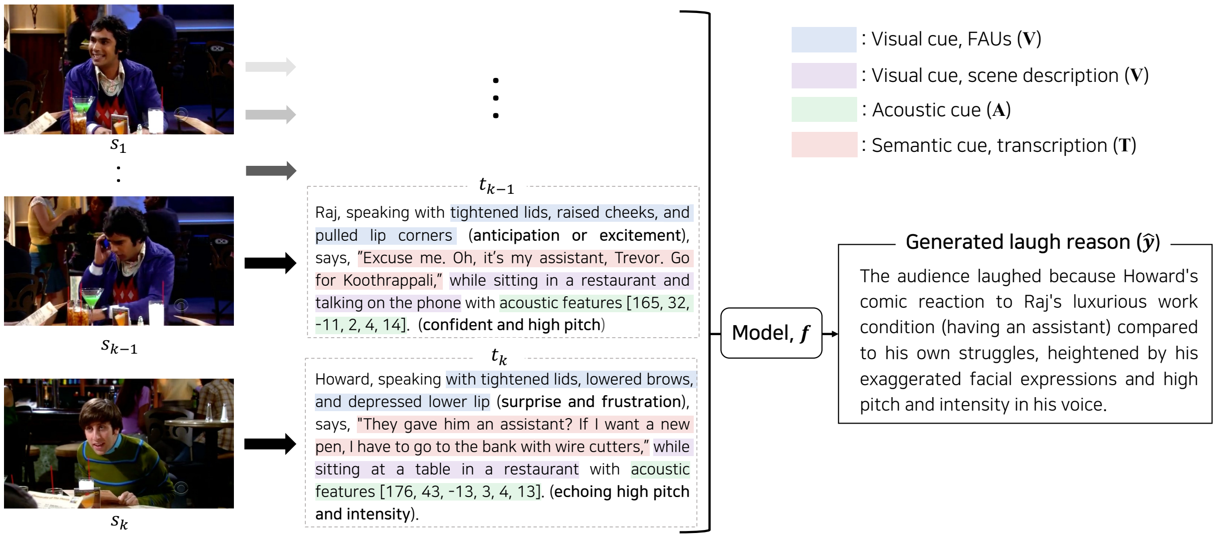


Figure 2: **Video Laugh Reasoning task and multimodal textual representation.** Each video clip (v) is trimmed into list of video segments (s_i), and each video segment is encoded into textual representation (t_i). The textual video representation consists of visual cues (V), acoustic cues from speech (A), and semantic cue (transcript, denoted as T). Then, we use LLM to generate why the audience laughs at the given video with the prompt. The bold text in parentheses on the t shows that LLM is semantically aware of the textual video representation.

Number of Video Clips	887
Number of Train/Val/Test	727 / 80 / 80
Number of Video Segments	4,434
Avg. number of Segments per clip (k)	4.4
Avg. duration of Video Clips	27.5 sec.
Avg. duration of Video Segments	6.2 sec.

Table 1: **Statistics of our dataset.** We split our dataset into train, validation, and test sets with the ratio of 8:1:1. Avg. denotes average.

average duration for TED talk clips is longer than *sitcoms*, given the protracted nature of talks.

Given that a single video clip often contains multiple instances of laughter, we focus on the last laugh in a clip for easier annotation. We only use video clips that meet the following filtering criteria, using a laugh detector (Gillick et al., 2021) to identify audience laughter instances. Our filtering criteria are: laughter should last at least 0.5 seconds, and be no more than 1 second interval between the video clip's last utterance and the onset of laughter. The latter criterion filters out the laughter events that are not related to the punchlines but are induced by something else. After this pre-processing, our final dataset comprises 484 sitcom and 403 TED talk video clips. Table 1 shows the statistics of our dataset.

Annotation for laughter reason. We employ human annotators from Amazon Mechanical Turk (AMT) to label videos with reasons for laughter. Given the inherently subjective nature of humor and the extensive variability in laughter triggers, constructing ground truth (GT) by free-form annotation is challenging. To mitigate these issues, we utilize the language model to generate candidates for laughter reasons, these candidates are subsequently presented to annotators with the corresponding video clip to choose the most appropriate explanation among them and refine it. If none of the candidates were suitable, we instruct them to write the reason in a free form.

After annotation, we verify all GT and manually refine it if it is not plausible for laughter reasons with video. This approach reduces the annotation workload by interacting LLM and humans, developing a more concise GT for this complex and subjective task. Finally, our dataset is formed as $\mathcal{D} = \{v, y\}$, where y is a GT explanation for laughter in the video clip v. See Appendix B for details about the human annotation process and the post-processing. Also, refer to Appendix F for the details about the AMT configuration.

3.3 Data Analysis

Which multimodal cue is important to infer the reason for laughter. We conduct a human evaluation to understand our dataset better. The annotators are requested to rank the multimodal cues in

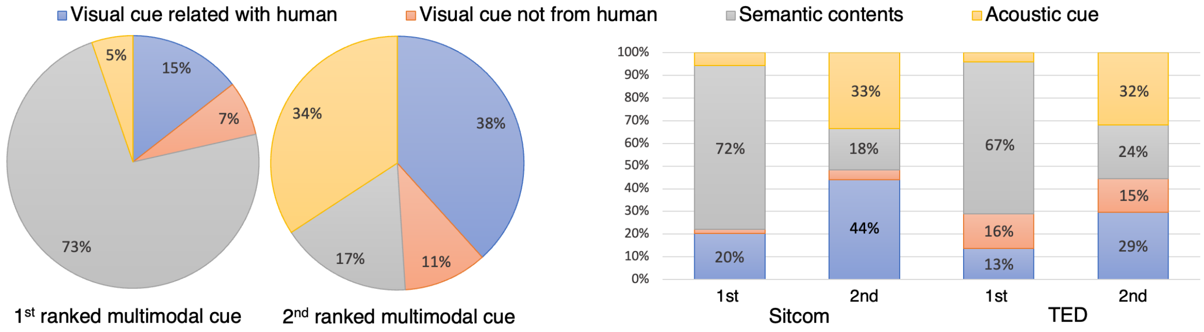


Figure 3: **Which multimodal cue is important to reason the laughter?** While semantic content is the most influential in causing laughter, the 2nd ranked modality cues are diverse, suggesting that multiple modality information can simultaneously influence laughter.

```
"prompt": {Reasoning task: you are to answer why the audience laughed given the video clip. The video clip from the {Sitcom}, titled {video title}, with multimodal information (Utterance, Facial Action Units, Video caption, Acoustic features(6 dimension; 1.mean of F0 contour, 2.var of F0 contour, 3. mean of energy contour, 4. var of energy contour, 5. jitter, 6. shimmer)) is given. The audience laughing moment is marked as (audience laughing) in certain utterance Explain why the audience laughed given the video clip, at most {30} words, starting with 'The audience laughed because '. Given video clip: {query}.}

"completion": {answer}
```

Figure 4: **Prompt for laugh reasoning experiments on GPT3.** The prompt is fed into GPT3 (Brown et al., 2020a) for fine-tuning, zero-shot learning, and in-context learning. For in-context learning, three random samples of prompt-answer pairs from the training set are given to GPT3. We manually change video types (sitcom or TED) and video title using the meta information of video clips. The query stands for multimodal textual representation m of the video clip. The length of the generated output is also variable, with a maximum of 30 words for sitcoms and 40 words for TED talks, considering each video type's characteristics.

perspective of which cues are related to laughter in the video. The rank annotation provides insight into which modality information is crucial for the cause of the laughter for each case.

For each video clip, we present annotators four choices: 1) visual cues from human; *e.g.*, facial expression and body gesture, 2) visual cues not from human; *e.g.*, backgrounds or images and props, 3) semantic contents; *i.e.*, transcription, and 4) acoustic cues; *e.g.*, speech tone or intensity. We ask them to choose two modality cues that are the most relevant for inducing laughter. The pie chart on the left in Figure 3 shows the modality importance statistics for our dataset. While the reason for laughter is primarily driven by semantic contents, the second most effective cue varies across different modalities, indicating that the various modalities in the video contribute to the reason for laughter. The bar chart on the right in Figure 3 shows the elements that induce laughter in two video types of our dataset. Notably, visual cues unrelated to humans, such as backgrounds or images, significantly trigger more laughter in TED than in sitcoms. TED videos often exhibit the speaker's presentation slides, making non-human visual cues more influential for eliciting laughter. Conversely, visual cues such as facial expressions and body gestures have a higher probability of causing laughter in sitcoms than in TED. This difference is because sitcoms mainly center around the characters' dialogues, so visual cues from human actors are more crucial. See Appendix C for additional data analysis.

Model	BLEU$_4$ (↑)	METEOR (↑)	ROUGE$_L$ (↑)	BERTScore (↑)	Win rate
Video model	0.226	0.236	0.398	0.427	24%
LLM + multimodal	**0.270**	**0.256**	**0.432**	**0.496**	**76%**

Table 2: **Comparison with video model.** We compare the video model trained on raw video and transcripts with LLM trained on multimodal textual representation. We use Video-LLaMA (Zhang et al., 2023) and LLaMA (Touvron et al., 2023) for video model and LLM, respectively.

4 Experiment

We split our dataset into 5 cross-validation splits except for the test set. We fine-tune two LLMs, GPT-3 (Brown et al., 2020a) and LLaMA (Touvron et al., 2023) with the training set and use the test set for evaluation.

Model	Num. of parameters	Modality	$BLEU_4$ (↑)	METEOR (↑)	$ROUGE_L$ (↑)	BERTScore (F1) (↑)
LLaMA (FT)	13B	T	0.250	0.245	0.432	0.493
		A+V+T	0.270	0.256	0.453	0.496
GPT-3 (zero-shot)	175B	T	0.126	0.155	0.313	0.389
		A+V+T	0.157	0.184	0.364	0.454
GPT-3 (3-shot)	175B	T	0.187	0.198	0.368	0.431
		A+V+T	0.232	0.230	0.413	0.476
GPT-3 (FT)	175B	T	0.230	0.243	0.429	0.488
		A+V+T	**0.279**	**0.267**	**0.475**	**0.523**

Table 3: **Evaluation on laugh reasoning with LLMs.** We evaluate whether the model can explain why the audience laughed. We fine-tune two LLMs, GPT-3 (Brown et al., 2020a) and LLaMA (Touvron et al., 2023) on our dataset, SMILE. We use GPT-3 for in-context (3 shots) and zero-shot experiments. Each modality cue in our dataset is denoted as Transcript (T), Audio (A), and Visual (V). FT denotes fine-tuning the model.

Implementation details. We use the official GPT-3 (Brown et al., 2020a), a non-free commercial version, as follows. We utilize the *davinci-text-002* model of GPT-3 (Brown et al., 2020a) for the zero-shot and in-context learning experiments. Examples of the prompts for both tasks are shown in Figure 4. The "prompt" provides the context of the task and the multimodal cues of the video, and "completion" provides the reason for the laughter. The zero-shot setup only takes "prompt" and generates the reason for the laughter, while the in-context learning setup is given with additional three randomly labeled samples from the training set as few-shot examples. More implementation details including LLaMA are in Appendix D.

Evaluation metrics. We utilize both quantitative metrics and human evaluation. We use metrics commonly employed for evaluating language generation tasks, including $BLEU_4$ (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), $ROUGE_L$ (Lin, 2004), and BERTscore (Zhang et al., 2019). For the human evaluation, we gather assessments from 3 crowd-workers per test sample by asking them to select their preferred explanation for laughter from a pair of options and take a majority vote to determine a winner. We calculate the average win rate (%) over the test set.

4.1 Comparison with video model

In addressing the laugh reasoning task, a direct method is to train a video model with raw video input. We compare the video model with our baseline, which utilizes LLM with multimodal textual representation. We fine-tune each model and conduct the quantitative and human evaluations (win rate), as shown in Table 2. The LLM-based baseline outperforms all metrics, indicating that our multimodal textual representation incorporates LLM's capacity to understand the reason for laughter in the video.

	A	B	A wins (%)	Fleiss'-κ
Q1	GPT-3 (A+V+T)	GPT-3 (T)	72.2	0.43
Q2	GPT-3 (FT)	GPT-3 (3-shot)	77.8	0.31
Q3	GPT-3 (FT)	LLaMA (FT)	56.6	0.49
Q4	Human	GPT-3 (FT)	66.2	0.42

Table 4: **Pairwise human evaluation.** Except for Q1, we use all modality (A+V+T) for training. We use Fleiss'-κ (Fleiss et al., 2013) for assessing the reliability of agreement. Q1-Q4 denote corresponding evaluation in § 4.2.

4.2 Evaluation

We analyze our baseline on laugh reasoning in various setups. We utilize both quantitative and human evaluation. Quantitative results are in Table 3, and the results of human agreements are in Table 4. Our evaluations aim to address four key questions.

Q1. Does multimodal information help for laugh reasoning? Yes, incorporating all modality cues for training enhances the performance of the laughter reasoning task compared to using transcripts alone (Table 3). The model trained with all modalities preferred in 72.2% of the test set compared to the transcript-only model as shown in Table 4. Furthermore, Fig. 5 (a) supports this, showing that the model trained with all modalities can effectively distinguish the reasons for laughter by utilizing multimodal information, whereas a transcript-only model only achieves a partial understanding.

Q2. Does the fine-tuning step help for a laugh reasoning? Yes, fine-tuned models outperform zero-shot/in-context models in both quantitative evaluation and human preference. It shows that our dataset nicely infuses the video laugh reasoning capacity to LLM.

"back in the and when i was growing up little girls were supposed to be kind and thoughtful"

Liza says "… fit into roles that were sort of shadowy really not quite clear what we were supposed to be (audience laughs)" with **cartoon of a boy and girl standing next to each other** with …

Why the audience laugh?

GPT-3 (FT) w/ A+V+T

The audience laughed because Liza Donnelly humorously described the shadowy roles of girls, which was **emphasized by the cartoon of a boy and girl** standing together.

GPT-3 (FT) w/ T

The audience laughed because Liza Donnelly humorously described how little girls were supposed to be kind and thoughtful, but their roles were unclear.

"There's a reciprocity clause. You get to pull the plug on him, too"

Sheldon speaking with **tightened lids, pulled lip corner, and raised cheeks** says "… Here's your I.D. card, your key and your lapel pin. Which Leonard was too cool to wear (audience laughs)", while sitting at a table …

GPT-3 (FT) w/ A+V+T

The audience laughed because Sheldon made a sarcastic comment about Leonard's coolness, which was **highlighted by his exaggerated facial expressions** and the audience's understanding of the characters' personalities.

GPT-3 (FT) w/ T

The audience laughed because the speaker made a sarcastic comment about Leonard's coolness, which was stressed by his failure to wear the lapel pin.

(a)

"Okay, I am officially unpacked. Thanks for helping me. Joey?"

"Oh my god, you almost gave me a heart attack!" (laughing)

GPT-3 (FT) w/ A+V+T

The audience laughed due to Chandler's surprised reaction and high-pitched response to a Joey's sudden appearance.

GT

The audience laughed because Chandler knew Joey was there and he was just acting surprised to a sudden appearance.

(b)

Figure 5: **Qualitative results on laugh reasoning.** For the examples in (a), GPT-3 (Brown et al., 2020a) fine-tuned on our dataset (denoted FT w/ A+V+T) understands the reasons for laughter by referencing multimodal cues. In contrast, the model fine-tuned using the transcript-only (denoted FT w/ T) manages to understand the reasons partially. The visual cues (scene description) are crucial for capturing "joey's sudden appearance" which is important to infer the reason for laughter in (b).

Q3. Do bigger models generate better reasons for laughter? Yes, GPT-3 (175B) surpasses LLaMA (13B) in both quantitative evaluation and human preference, as shown in Table 3 and 4.

Q4. Does the model explain the reason for laughter as well as humans? No, the human-annotated laughter reasons are preferred by 66.2% than those generated by fine-tuned GPT-3 (our best model) as shown in Q4 of Table 4. Figure 5 (b) provides an example illustrating the comparison between human-annotated reasons (GT) and generated reason for laughter. In this sample, all crowd workers prefer GT because the model struggles to distinguish the subtle difference between surprise and posed surprise, while the human-annotated reason successfully captures it.

In summary, for the laugh reasoning task, multimodal information, a large model, and infusing reasoning capacity with our dataset are important. While the trained model does not surpass human capabilities, the use of LLM with multimodal textual representation enables us to generate plausible explanations about the reason for laughter in videos. See Appendix E for additional experiments.

Model	MUStARD Acc. (%) (↑)	UR-FUNNY Acc. (%) (↑)
TFN (Zadeh et al., 2017)	68.6	64.7
CMFN (Hasan et al., 2019)	70.0	65.2
MISA (Hazarika et al., 2020)	66.1	70.6
BBFN (Han et al., 2021)	71.4	71.7
MUStARD++ (Ray et al., 2022)	74.2	-
MAG-XLNet (Rahman et al., 2020)	74.7	72.4
MuLoT (Pramanick et al., 2022)	76.8	73.9
Ours (w/ LLaMA)	77.5	75.1
Ours (w/ GPT-3)	**79.0**	**77.9**

Table 5: **Evaluation results of the humor & sarcasm detection task.** All models use text, visual, and acoustic information from videos for training.

Figure 6: **Examples of in-the-wild videos.** We try to generate explanations for the laughter in the videos from standup comedy and intimate conversation. The results show that we can generate valid explanations for laughter in other videos.

5 Discussion

In this section, we discuss the scalability of utilizing large language models with textual video representation by conducting evaluations on other tasks and on in-the-wild videos.

5.1 Evaluation on other tasks

Apart from laugh reasoning, we conduct humor detection and sarcasm detection tasks, which classify given video contains humor (sarcasm) or not (i.e., binary classification). We use UR-FUNNY (Hasan et al., 2019) and MUStARD (Castro et al., 2019), which are representative benchmarks for these tasks. We cast the original binary classification problem as a text generation problem to integrate into our system. Formally, we can define the task as, $\hat{b} = f(\mathcal{P}, \{t_1, t_2, ..., t_k\})$, where $\hat{b}$ denote predicted binary class in text format ("Yes" or "No"), and $\mathcal{P}$ is prompt for instructing LLMs about the task and input representation.

We follow the same train/test split, and evaluation procedure as in the benchmark for measuring the accuracy of each detection task. We use LLaMA and GPT-3 for training with textual representation converted from the video in the training set of each benchmark dataset. Table 5 shows that our method achieves strong performance[4] on both tasks. This experiment highlights the scalability of utilizing LLMs with textual representation in various video understanding tasks.

[4]We do not compare with FunnyNet (Liu et al., 2022) as they use an additional large-scale dataset for training.

5.2 Evaluation on the in-the-wild videos

We extend our laughter reasoning to in-the-wild videos, encompassing different video types and laughter contexts compared to our dataset. First, we evaluate our approach on a video clip from a stand-up comedy, which has similar audience laughter patterns to those in our dataset. We convert the video into a textual representation and infer the reason for the audience laughing. Figure 6 shows that the model can generate a plausible explanation for the reason for laughter in stand-up comedy.

Next, we test on a video clip featuring an intimate conversation between a married couple. In this case, the laughter originates from the speakers themselves rather than from the audience. As this does not belong to the comedic genre but rather a sincere conversation between two people, it is more likely that non humor-based laughter, such as nervous or social laughter, may occur. Figure 6 shows that the model can also understand the nervous laughter used to alleviate tension or awkwardness in the situation.

6 Conclusion

In this paper, we aim to understand the reason behind laughter by introducing *Laugh Reasoning* task, accompanied with SMILE dataset. While the model did not surpass human capabilities, we show that the model can generate plausible explanations about laughter reason, underlining that multimodal cues in our dataset nicely infuse the laugh reasoning capacity to the model. We also show the results applied to other tasks and other types of video, hinting at the scalability of utilizing LLM with multimodal textual representation.

Limitation & future direction. Our LLM-based baseline serves as the initial method for laugh reasoning task and has a margin to improve. For the multimodal textual representation, as it is a primitive form for capturing human social interaction in the video, we can enhance it with diverse attributes such as gesture, eye gaze, and relationship or use other representations such as scene graph. Our work mainly focuses on audience laughter as the first stepping stone toward understanding laughter due to its distinct and cohesive signal, while there are diverse mechanisms behind laughter. Recognizing this, enriching our work with diverse video types like vlogs, movies, and talk shows is a promising direction to capture a broader range of laughter, as we show the possibility in § 5.2.

Potential application & broader impact. Our work can be regarded as a stepping stone toward developing socially intelligent agents that understand and appropriately create non-verbal cues, such as laughter, playing a crucial role in building rapport, expressing emotions, and creating deep emotional exchanges (Tickle-Degnen and Rosenthal, 1990; Argyle, 1972). Such advancement moves us beyond the capabilities of current dialogue agents, e.g., ChatGPT or Alexa, which mostly focus on verbal signals. Incorporating 3D talking head methods (Sung-Bin et al., 2024; Zhao et al., 2024) could offer the way agents are visualized, enabling more expressive and multimodal interactions with users.

Acknowledgement. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No.2022-0-00290, Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense and No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities and No.2021-0-02068, Artificial Intelligence Innovation Hub) and NCSOFT.

References

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.

Issa Annamoradnejad and Gohar Zoghi. 2020. Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*.

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

Mahadev L Apte. 1985. *Humor and laughter: An anthropological approach*. Cornell university press.

Michael Argyle. 1972. Non-verbal communication in human social interaction. *Non-verbal communication*, 2(1).

Tomas Arias-Vergara, Juan Camilo Vásquez-Correa, and Juan Rafael Orozco-Arroyave. 2017. Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech. *Cognitive Computation*.

Salvatore Attardo. 2008. A primer for the linguistics of humor. *The primer of humor research*.

William Sims Bainbridge, Edward E Brent, Kathleen M Carley, David R Heise, Michael W Macy, Barry Markovsky, and John Skvoretz. 1994. Artificial social intelligence. *Annual review of sociology*.

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.

Dario Bertero and Pascale Fung. 2016. Deep learning of audio and language features for humor prediction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 496–501, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Billig. 2005. *Laughter and ridicule: Towards a social critique of humour*. Sage.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*.

Najim Dehak, Pierre Dumouchel, and Patrick Kenny. 2007. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*.

Stéphane Dupont, Hüseyin Çakmak, Will Curran, Thierry Dutoit, Jennifer Hofmann, Gary McKeown, Olivier Pietquin, Tracey Platt, Willibald Ruch, and Jérôme Urbain. 2016. Laughter research: a review of the ilhaire project. *Toward Robotic Socially Believable Behaving Systems-Volume I: Modeling Emotions*, pages 147–181.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Sigmund Freud. 1960. *Jokes and their relation to the unconscious*. WW Norton & Company.

William F Fry. 2011. *Sweet madness: A study of humor*, volume 1. Transaction Publishers.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman. 2021. Robust laughter detection in noisy environments. In *Conference of the International Speech Communication Association (Interspeech)*.

David Greatbatch and Timothy Clark. 2003. Displaying group cohesiveness: Humour and laughter in the public lectures of management gurus. *Human relations*.

Charles R Gruner. 1978. *Understanding laughter: The workings of wit & humor*. Burnham Incorporated Pub.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. 2023. Champagne: Learning real-world conversation from large-scale web videos. *IEEE International Conference on Computer Vision (ICCV)*.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *ICMI*.

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.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM International Conference on Multimedia (MM)*.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor 'understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2022. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Dayoon Ko, Sangho Lee, and Gunhee Kim. 2023. Can language models laugh at youtube short-form videos? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2897–2916.

Jin Joo Lee et al. 2017. *A Bayesian theory of mind approach to nonverbal communication for human-robot interactions: a computational formulation of intentional inference and belief manipulation*. Ph.D. thesis, Massachusetts Institute of Technology.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*.

Zhisong Liu, Robin Courant, and Vicky Kalogeiton. 2022. Funnynet: Audiovisual learning of funny moments in videos. In *Asia Conference on Computer Vision (ACCV)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations (ICLR)*.

Rod A Martin and Thomas Ford. 2018. *The psychology of humor: An integrative approach*. Academic press.

Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of research in personality*.

Paul McDonald. 2013. *The philosophy of humour*. Humanities-Ebooks.

Gary McKeown, Roddy Cowie, Will Curran, Willibald Ruch, and Ellen Douglas-Cowie. 2012. Ilhaire laughter database. In *Proceedings of 4th International*

Workshop on Corpora for Research on Emotion, Sentiment & Social Signals, LREC, pages 32–35.

Harvey Mindess. 2017. *Laughter and liberation*. Routledge.

Anton Nijholt, Andreea I Niculescu, Alessandro Valitutti, and Rafael E Banchs. 2017. Humor in human-computer interaction: a short survey. *Adjunct Proceedings of INTERACT*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Robert R Provine. 2001. *Laughter: A scientific investigation*. Penguin.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. 2022. A multimodal corpus for emotion recognition in sarcasm. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6992–7003, Marseille, France. European Language Resources Association.

Sophie K Scott, Nadine Lavan, Sinead Chen, and Carolyn McGettigan. 2014. The social life of laughter. *Trends in cognitive sciences*.

David Stauffer. 1999. Let the good times roll: Building a fun culture. *Harvard Management Update*.

Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. 2024. Laughtalk: Expressive 3d talking head generation with laughter. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6404–6413.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *ACM International Conference on Multimedia (MM)*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jérôme Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner. 2010. The AVLaughterCycle database. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Robert A Vartabedian and Laurel Klinger Vartabedian. 1993. Humor in the workplace: A communication challenge.

Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Learning common sense through visual abstraction. In *IEEE International Conference on Computer Vision (ICCV)*.

Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning (ICML)*. PMLR.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022b. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022c. Language models with image descriptors are strong few-shot video-language learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Melissa B Wanzer, Ann B Frymier, and Jeffrey Irwin. 2010. An explanation of the relationship between instructor humor and student learning: Instructional humor processing theory. *Communication education*.

Caleb Warren, Adam Barsky, and A Peter McGraw. 2021. What makes things funny? an integrative review of the antecedents of laughter and amusement. *Personality and Social Psychology Review*.

Orion Weller and Kevin Seppi. 2020. The rJokes dataset: a large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France. European Language Resources Association.

Jessica Williams, Stephen M Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li Yao, Yan Wan, Hongjie Ni, and Bugao Xu. 2021. Action unit classification for facial expression recognition using active learning and svm. *Multimedia Tools and Applications*.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, San Diego, California. Association for Computational Linguistics.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations (ICLR)*.

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

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. 2017. S3fd: Single shot scale-invariant face detector. In *IEEE International Conference on Computer Vision (ICCV)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Qingcheng Zhao, Pengyu Long, Qixuan Zhang, Dafei Qin, Han Liang, Longwen Zhang, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2024. Media2face: Co-speech facial animation generation with multi-modality guidance.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Multimodal Textaul Representation

In this section, we explain how to convert video into multimodal textual representation. Videos are multimodal, which include visual, acoustic, and semantic cues (i.e., transcription). We encode video clips into textual representation, embracing their multimodal information, so that we can leverage the pre-trained knowledge of LLMs while exploiting multimodal inputs in our baselines. First, starting with a video clip, we build a list of video segments by trimming the clip based on the utterances. The definition of the utterance varies upon to the source of the video: for TED talks, each sentence is defined as an utterance, since TED talk usually has a single speaker. If the utterance is too short (2 seconds or less), we concatenate adjacent utterances into one. For sitcoms, we define consecutive sentences from the same speaker as an utterance.

Visual cues. We compose visual cues with facial expressions and scene descriptions to perceive human-specific and scene-wide contextual information. Specifically, to process human-specific information, we utilize the active speaker detection algorithm (Tao et al., 2021) and face detector (Zhang et al., 2017) to crop the face of the speaking person in each video segment. This process effectively identifies the active speaker, especially for sitcoms where many people appear in a single scene, allowing to align visual features with utterances.[5] For facial expression description, we extract 14 facial action units (FAUs) (Yao et al., 2021)[6] from each frame in the video segment with 10 frames per second (FPS).

Then, we accumulate them and take the three most dominant units. For scene-wide contextual cues, we use the video captioning (Wang et al., 2022b) to extract scene description. The scene description provides high-level context for the visual cues including the surrounding objects and background that interact with the speaker.

Acoustic cues. We extract the mean and the variance of pitch, intensity, jitter and shimmer as acoustic features from speech utterance using off-the-shelf speech processing models (Arias-Vergara et al., 2017; Dehak et al., 2007). Since the extracted values are real numbers, we initially try to convert them to a linguistic format with certain criteria (*e.g.*, map to "high pitch" if the mean pitch value is greater than 200). However, it is challenging to set an objective criterion that considers various factors, including the speaker's gender, context, and identity. Instead of putting real numbers into text, we use themselves as acoustic features by giving a description of them as a prompt to LLMs, leveraging their knowledge on understanding numerical number (Brown et al., 2020b; Liu et al., 2023; Jiang et al., 2020; Wallace et al., 2019) (See bold text in parentheses on the t in Figure 2).

[5]We provide these face-cropped video segments in our dataset.
[6]We use https://github.com/CVI-SZU/ME-GraphAU to extract FAUs.

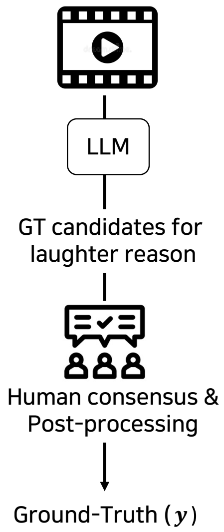


Figure 7: **Annotation pipeline for laughter reason.**

B Annotation for Laughter Reason

We elaborate the procedure for obtaining laughter reason consensus (ground-truth; GT) by utilizing large language models' general knowledge and incorporating it into human consensus. This procedure consists of three steps: (1) build GT candidates, (2) human annotation, and (3) post-processing (See Figure 7).

For (1) building GT consensuses, we utilize the large language model (GPT-3.5 (Ouyang et al., 2022)) with multimodal textual representation t to generate two candidates for the laughter reason. We manually pre-process these candidates if they are invalid or have incorrect sentence structure (See Figure 8).

For (2) human annotation, the processed GT candidates are subsequently presented to annotators from Amazon mechanical turk (AMT) with the corresponding video clip. The annotators are asked to choose the most appropriate explanation among them. If the annotators judge that no candidates are appropriate, we instruct the annotators to write

Initial GT candidate (*can not find the laughter reason*)

`There is no clear indication of audience laughter in the given context.`

Pre-processed GT candidate

`The audience laughed because Sophie Scott is playing on the screen where they see a man in his underwear trying to endure the coldness of the ice and ended up giving up after a short period of time.`

Initial GT candidate (*incorrect instructed sentence structure*)

`There is one instance where speaker joked about diagnosing genetic conditions in their kitchen sinks or doing at-home paternity testing. The audience laughed because it was perceived as a humorous remark.`

Pre-processed GT candidate

`The audience laughed because the speaker mentioned doing at-home paternity testing and diagnosing genetic conditions in the kitchen sink, which is a humorous and absurd concept`

Figure 8: **Examples of the pre-processing of GT candidates.** In the example above, GPT3.5 fails to infer the reason for the laughter given a multimodal textual representation of the video clip. We handle this by utilizing GPT4 to generate reasons for laughter from the same input. In the example below, the sentence structure does not start with “The audience laughed because”, which is the structure we want. In this case, we manually revise it for consistent sentence structure.

Given GT candidates

1. `The audience laughed because Sophie’s statement “he’s not laughing yet” was followed by the video showing a man holding a frisbee, creating a humorous contrast between statement and visual context.`
2. `The audience laughed because the speaker made an ironic remark about a man not yet laughing, as he held a frisbee, creating humorous contrast.`

Refined GT written by annotators in free-form

`The audience laughed because Sophie is playing on the screen where they see a man in his underwear trying to endure the coldness of the ice and ended up giving up after a short period of time, creating a humorous contrast between the statement and the visual context.`

Given GT candidates

1. `The audience laughed because Chandler made a joke about his testicles possibly being in the box, which was unexpected and inappropriate, causing amusement.`
2. `The audience laughed because Chandler made a suggestive joke about his "testicles" being in a gift box, which was a humorous and unexpected innuendo.`

Refined GT written by annotators in free-form

`The audience laughed because Chandler's joke about his testicles being in the box is because he knows so much about ribbon types which makes him seem feminine.`

Figure 9: **Examples of the correction of laughter reason by annotators.** All given GT candidates are passed to the annotators after pre-processing. The free-form responses capture additional visual details (above) and provide a context of why saying “testicles” evokes laughter (below).

or refine the reason in free form. After annotation, the candidate with the most votes is selected as the GT. If at least one annotator provided the reason for laughter in free-form, we manually checked their validity and reflected them into GT. Figure 9 shows that free-form responses capture additional visual details and provide an understanding of why certain words elicit laughter. See Appendix F for details about AMT.

For (3) post-processing, we additionally verify all GTs for laughter reasons and manually refine it if it is not plausible for laughter reasons with video or has repetitive phrases that might induce spurious correlation. To mitigate this, we replace repeated phrases with synonyms, which are randomized among multiple synonyms. For example, one of the repetitive phrases “unexpected and humorous”, is randomly replaced with synonyms such as “astonishing and laughable”, or “hilarious”. As another correction, even with the best efforts of human annotators, some reasons are not perfectly matched with the video. Figure 10 shows the post-processing that corrects these kinds of errors.

Annotation quality control. We use qualification criteria to ensure the annotation quality. We allow annotators from (AU, CA, NZ, GB, US), which represent the English-speaking countries.[7] Additionally, we only allow experienced annotators who are with 10K approved previous HITs and a minimum acceptance rate of 97% on their previous HITs. We pay each annotator 0.3 USD($) per accepted HIT.

C Data Analysis

We further conduct a human evaluation to understand our dataset better. Given the video clip, the

[7]This is because all the video clips in our dataset are in English.

GT (*repetitive phrase*)

`The audience laughed because the doctor's diagnosis of Sheldon's inflamed larynx was exaggerated, and the use of the phrase "I've never seen anything like it" was unexpected and humorous.`

Post-processed GT

`The audience laughed because the doctor's diagnosis of Sheldon's inflamed larynx was exaggerated, and the use of the phrase "I've never seen anything like it" was astonishing and laughable.`

GT (*not plausible for laughter reason within video*)

`The audience laughed due to Chandler make noise with high-pitched tone and exaggerated facial expression to Rachel.`

Post-processed GT

`The audience laughed because Chandler make fun of Rachel with her appearance with high-pitched tone and exaggerated facial expression`

Figure 10: **Examples of the post-processing on GT.** The example above shows replacing a repetitive phrase with a synonym. The example below shows how we rectify GT when the reason for laughter does not align with the video context.

annotators are requested to determine the laugh. The laugh type annotation explains the distinct characteristics of laughter in TED and sitcoms.

We consider two laugh types: 1) *Release-Triggered Laughter* (Freud, 1960; Fry, 2011; Mindess, 2017) that results from the alleviating tension amidst constraints such as awkward or complex situation and 2) *Hostility-Triggered Laughter* (Gruner, 1978; Billig, 2005) that arises from claiming superiority over someone or something, based on “great families” of theories of humor (Attardo, 2008), and ask annotator to determine which one is more appropriate for laughter in video.[8]

Statistics in Figure 11 suggest that sitcoms and TED talks are dominated by different types of laughter, suggesting that the nature of laughter varies by video type. Specifically, the major laugh type in sitcoms is closer to the hostility-induced laughter, and we postulate that sitcoms are typically designed to be entertaining, focusing on humorous situations, witty dialogue, and comedic conflicts among characters. On the other hand, TED talks are dominated by release-triggered laughter. We hypothesize that the talks aim to captivate and engage the audience by releasing constraints and unexpected revelations, creating a dynamic and thought-provoking experience. This type of humor helps maintain interest, and breaks the monotony (Wanzer et al., 2010). By merging these two heterogeneous video types, we can cover a wider range of reasons behind the audience’s laughter.

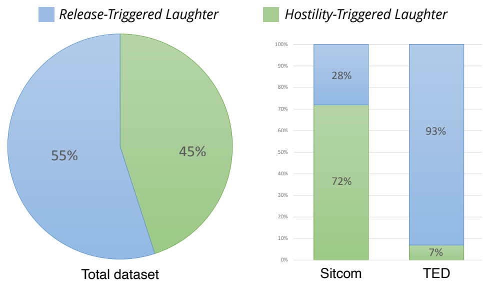


Figure 11: **Laughter types in our dataset.** Sitcoms tend to have more hostility-triggered laughter, while TED talks have more released-triggered laughter.

D Implementation details

GPT3 fine-tuning. We utilize the OpenAI fine-tuning API and fine-tune *davinci*. The prompt for fine-tuning is the same as the aforementioned experiments. We follow the fine-tuning scheme provided on the OpenAI webpage.[9]

LLaMA fine-tuning. LLaMA is LLM, an open-source model for research. We fine-tune the full parameters of LLaMA for 5 epochs. We utilize 4 A100 (80GB) for distributed fine-tuning with batch size 4 per device and a learning rate 1e-4. We also leverage fp16 mixed precision.

Video-LLaMA fine-tuning. We use Video-LLaMA which consists of pre-trained Blip2, Vicuna-13B, and Imagebind-huge. We train audio, video Q-former, and projection layers while other parameters are frozen. We utilize 8 A100 (80GB) for distributed fine-tuning with batch size 1 per device and an initial learning rate (3e-5), and weight decay (0.05) for 10 epochs. We also leverage mixed

[8]During annotation, we provided full descriptions of the concepts of the laughter types, rather than using the terms.

[9]https://platform.openai.com/docs/guides/fine-tuning; OpenAI has not opened the details of the API’s fine-tuning mechanisms, which is currently hidden.

```
"prompt": {Humor detection task: given video clip from the {TED}, titled {video title}, with multimodal
information (Utterance, Facial Action Units, Video caption, Acoustic features(6 dimension; 1.mean of F0 contour,
2.var of F0 contour, 3. mean of energy contour, 4. var of energy contour, 5. jitter, 6. shimmer)) is given. The
audience laughing moment is marked as (audience laughing) in certain utterance. Given video clip: {query}, Is
the video contain humor?, answer in yes or no (binary classification)
"completion": {answer}

"prompt": {Sarcasm detection task: given video clip from the {sitcom}, titled {video title}, with multimodal
information (Utterance, Facial Action Units, Video caption, Acoustic features(6 dimension; 1.mean of F0 contour,
2.var of F0 contour, 3. mean of energy contour, 4. var of energy contour, 5. jitter, 6. shimmer)) is given. The
audience laughing moment is marked as (audience laughing) in certain utterance. Given video clip: {query}, Is
the video contain sarcasm?, answer in yes or no (binary classification)}
"completion": {answer}
```

Figure 12: **Prompt for humor and sarcasm detection.** We manually change video types (sitcom or TED) and video title (such as Patrick Chappatte (2010 Global) or BBT) using the meta information of video clips. The query stands for multimodal textual representation m of the video clip. Answer denote label (yes or no) from UR-FUNNY (Hasan et al., 2019) and MUStARD dataset (Castro et al., 2019).

Test dataset	Train dataset	Modality	BLEU$_4$ (↑)	METEOR (↑)	ROUGE$_L$ (↑)	BERTScore (F1) (↑)
SMILE$_{\text{Sitcom}}$	SMILE$_{\text{Sitcom}}$	T	0.214	0.248	0.429	0.489
		A+V+T	0.290	0.288	0.485	0.548
	SMILE	T	0.241	0.252	0.446	0.510
		A+V+T	**0.298**	**0.289**	**0.499**	**0.555**
SMILE$_{\text{TED}}$	SMILE$_{\text{TED}}$	T	0.260	0.241	0.432	0.459
		A+V+T	**0.279**	**0.260**	**0.454**	0.457
	SMILE	T	0.249	0.245	0.423	0.454
		A+V+T	0.273	0.247	0.438	**0.468**

(a) Video type-wise evaluation

Test dataset	Train dataset	Modality	BLEU$_4$ (↑)	METEOR (↑)	ROUGE$_L$ (↑)	BERTScore (F1) (↑)
SMILE$_{\text{Sitcom}}$	SMILE$_{\text{TED}}$	A+V+T	0.161	0.254	0.390	0.407
	SMILE$_{\text{Sitcom}}$	A+V+T	0.290	0.288	0.485	0.548
SMILE$_{\text{TED}}$	SMILE$_{\text{Sitcom}}$	A+V+T	0.153	0.193	0.369	0.449
	SMILE$_{\text{TED}}$	A+V+T	0.279	0.260	0.454	0.457

(B) Cross-dataset evaluation

Table 6: **Analysis on video types.** In (a), we conduct the video type-wise evaluation as the dominant laughter type differs along the video type. In (b), we evaluate the model by testing on the different video types, *i.e.*, cross-dataset.

Test	A	B	A wins (%)	Fleiss'-κ
TED	GPT-3 (SMILE)	GPT-3 (TED)	66.2	0.40
Sitcom	GPT-3 (SMILE)	GPT-3 (sitcom)	61.4	0.63

Table 7: **Pairwise human evaluation.** We compare the model trained with the whole dataset (SMILE) with a subset (TED, sitcom) and evaluate them with the test set of each subset.

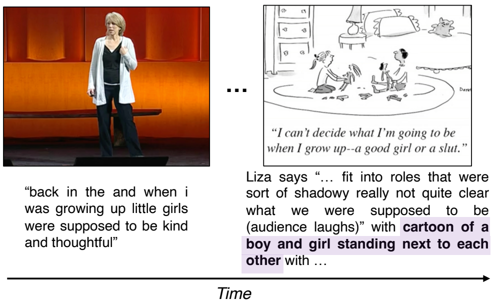
precision that uses fp16 for multiplication and fp32 for addition.

Detection. For the sarcasm (Castro et al., 2019) and humor detection (Hasan et al., 2019) tasks, we finetune LLaMA-13B (Touvron et al., 2023) and GPT-3 (Brown et al., 2020a) with our multimodal textual representation. GPT-3 finetuning is as same as described for the laugh reasoning task. For LLaMA-13B, we follow the fine-tuning script on Vicuna (Chiang et al., 2023)[10]. Examples of the prompts for both tasks that cast classification task to generation task are shown in Figure 12. We use four A100 (80GB) for each training. We follow Vicuna's default LLaMA fine-tuning hyperparameters except for setting the per-device batch size to 3 and the number of training epochs to 20.

E Additional Experiments

Evaluation by video types. The type of laughter varies depending on the source of the video, as shown in Figure 11. To explore this further, we evaluate each video type independently. Instead of fine-tuning GPT3 on the entire SMILE dataset, we separately fine-tune the models on subsets of

[10]https://github.com/lm-sys/FastChat



What is the background color of the cartoon?

VideoLaMMA

`The background color of the cartoon is white.`

VideoChat

`I think it is white.`

Why the audience laugh?

VideoLaMMA

`The audience laughed because the speaker made a humorous remark about women and power.`

VideoChat

`The audience laughed because the cartoon is funny.`

Ours

`The audience laughed because Liza Donnelly humorously described the shadowy roles of girls, which was emphasized by the cartoon of a boy and girl standing together.`

Figure 13: **Examples of laugh reasoning on recent video language models.** While recent video language models, such as Video-LLaMA (Zhang et al., 2023) and VideoChat (Li et al., 2023), can respond to the general question in the video, they struggle to plausibly explain the reason for the laughter in the video.

the dataset, namely SMILE$_{\text{Sitcom}}$ and SMILE$_{\text{TED}}$. As summarized in Table 6 (a), even when models are independently fine-tuned to different video types, their performance is comparable to that of the model trained on the SMILE dataset. Interestingly, in the human evaluation, the model trained on whole data (SMILE) is preferred over the model trained on each video type. This suggests that our dataset, SMILE, covers the diverse laughing characteristics to lead GPT3 to learn generalized laughter reasons across different types of videos.

However, we observe that testing the model across different video types, *e.g.*, training on SMILE$_{\text{Sitcom}}$ and testing on SMILE$_{\text{TED}}$, results in a significant performance drop, as shown in Table 6 (b). We speculate that this is due to differences in laughter types presented in each source video. This supports the idea that combining these two heterogeneous video types could help the model learn to understand a broader range of reasons behind audience laughter.

Video language model. While the previous methods (Zellers et al., 2019; Zadeh et al., 2019) have aimed to learn and reason about social interactions from visual data, they formulate the task in multiple-choice setups. By virtue of the advance of large language models, recent work has suggested multimodal models capable of generating natural language responses to questions about a video, rather than outputting a multiple-choice answer. In this context, we examine if these models can exhibit the capability to reason behind laughter in a given video. We feed the same video from Figure 5 into recent video-language (VL) models, Video-LLaMA (Zhang et al., 2023)[11] and VideoChat (Li et al., 2023)[12], and showcase their generated reasoning in Figure 13. While these models can respond to general questions about the video, they struggle to reason about moments of laughter. Unlike existing multimodal reasoning work, we contribute a new perspective to multimodal reasoning, aiming to understand and reason about an important social signal, laughter.

F Human annotation from Amazon Mechanical Turk

Figure 14 shows our interface and instructions for the annotators working on Amazon Mechanical Turk (AMT). We define a questionnaire per video clip as a Human Intelligence Task (HIT). We ask AMT annotators three questions in a HIT, 1) laughter reason, 2) laugh type, and 3) the multimodal cues in perspective of which cues are related to laughter in the video. The first question is for obtaining GT annotations for laughter reasons and pairwise human evaluation in § 4. The second and third questions are for the data analysis purpose, which provides further understanding of our dataset (See § 3.3 in the main paper and Appendix C).

[11]https://github.com/DAMO-NLP-SG/Video-LLaMA
[12]https://github.com/OpenGVLab/Ask-Anything

Instructions (click to expand)

Q.1 Pick the one that better explains why the audience laughs at the end of the video.

The audience laughed because the speaker whimsically juxtaposed the grandiose Washington memorials with London's minimalist tribute to David Lloyd George, which boasts nothing more than his name. It was an unexpected, comically stark contrast that tickled the audience's funny bone.

○ *choice_1*

The audience laughed because the speaker contrasted the elaborate memorials in Washington with the simple one in London, where the monument to David Lloyd George contained only his name.

○ *choice_2*

If the given choices do not provide a good explanation, please improve it or write a new one:

Q2 Pick the type of the laugh.

○

Hostility: This laughter often arises from the situations, where humor frequently emerges from comical conflicts, misunderstandings, or characters outwitting each other.

○

Release: Speakers often introduce humor to alleviate tension amidst complex topics or awkward situation.

Q3-1 Pick the information of video that most relevant for audience laughs.

○ *Visual: Facial expressions and body language*

○ *Visual: All visual cues that are not from human, including background, props, images, etc.*

○ *Semantic: The conversation contents (transcriptions)*

○ *Audio: Speech tones, intensity, speed*

Q3-2 Pick the information of video that second most relevant for audience laughs.

○ *Visual: Facial expressions and body language*

○ *Visual: All visual cues that are not from human, including background, props, images, etc.*

○ *Semantic: The conversation contents (transcriptions)*

○ *Audio: Speech tones, intensity, speed*

Submit

Instructions (click to expand)

In this HIT, you watch a short video clip from BigBang Theory, Friends, and TED, then assess why the audience laugh at the end of the video. The video may contain multiple audience laughs, but you **only need to deal with the last audience laughs caused by the last line or situation in the video.**

Q1. Pick the one that better explains why the audience laughs at the end of the video.

- You get two choices about the reason for laughing. Assess which one is more appropriate.
- *If there is no laughter in the video, assume there is after the last utterance (situation) and evaluate why*
- *If given all the choices doesn't provide a good explanation of why, please improve it or write a new one.*
- *Answer must start with "The audience laughed because ~"*

Q2. Pick the type of the laugh.

- *You get two choices about the types for laughing. Assess which one is more appropriate.*
  - *1) Hostility-Induced Laughter: This form of laughter often arises from the situations, where humor frequently emerges from comical conflicts, misunderstandings, or characters outwitting each other.*
  - *2) Release-Triggered Laughter: Speakers often introduce humor to alleviate tension amidst complex topics or awkward situation.*

Q3-1. Pick the information of video that most relevant for audience laughs.

- *You get four choices (information types); first pick the most relevant and help you to reason for laughing.*
  - *1) Visual: Facial expressions and body language.*
  - *2) Visual: All visual cues that are not from human, including background, props, images, etc.*
  - *3) Semantic: The conversation contents (transcriptions).*
  - *4) Audio: Speech tones, intensity, speed.*

Q3-2. Pick the information of video that second most relevant for audience laughs.

- *pick the second most relevant and help you to reason for laughing.*
- *Note: You must check different information in Q2-1, Q2-2*
  - *1) Visual: Facial expressions and body language.*
  - *2) Visual: All visual cues that are not from human, including background, props, images, etc.*
  - *3) Semantic: The conversation contents (transcriptions).*
  - *4) Audio: Speech tones, intensity, speed.*

Figure 14: **Examples of the AMT interface (left) and instructions (right) that the annotators worked on.** The annotators are asked to watch the video clip and answer the three questions. The third question is split into two parts. We put the instructions at the top of the interface to emphasize how the annotators should answer each question.