

MUStReason: A Benchmark for Diagnosing Pragmatic Reasoning in Video-LMs for Multimodal Sarcasm Detection.

Anisha Saha^{1, 2}, Varsha Suresh², Timothy Hospedales^{3, 4}, Vera Demberg^{1, 2}

¹Max Planck Institute for Informatics, Saarland Informatics Campus, ²Saarland University,

³The University of Edinburgh, ⁴Samsung AI Center, Cambridge

ansaha@mpi-inf.mpg.de, vsuresh@lst.uni-saarland.de,

t.hospedales@ed.ac.uk, vera@lst.uni-saarland.de

Abstract

Sarcasm is a specific type of irony which involves discerning *what is said* from *what is meant*. Detecting sarcasm depends not only on the literal content of an utterance but also on non-verbal cues such as speaker's tonality, facial expressions and conversational context. However, current multimodal models struggle with complex tasks like sarcasm detection, which require identifying relevant cues across modalities and pragmatically reasoning over them to infer the speaker's intention. To explore these limitations in VideoLMs, we introduce **MUStReason**, a diagnostic benchmark enriched with annotations of modality-specific relevant cues and underlying reasoning steps to identify sarcastic intent. In addition to benchmarking sarcasm classification performance in VideoLMs, using MUStReason we quantitatively and qualitatively evaluate the generated reasoning by disentangling the problem into perception and reasoning and aim to pinpoint the current gaps in these VideoLMs. Furthermore, to facilitate structured pragmatic reasoning, we propose PragCoT, a framework that steers VideoLMs to focus on implied intentions over literal meaning, a property core to detecting sarcasm. Code and dataset are available at <https://github.com/anisha0325/MUStReason>.

Keywords: Multimodal Sarcasm Detection, Video-Language Models, Pragmatic Reasoning

1. Introduction

Conversations in day-to-day life often involve the use of rhetorical devices like humor, irony, and sarcasm. These forms of expression have an underlying intent that is opposite to their literal meaning. Some forms of sarcasm, such as “*Never? Isn't that usually when you go to the gym?*”, is explicitly conveyed through the speaker's utterance and can be directly identified from the text. In contrast, others like “*It was her, right?*” rely on additional contextual or multimodal cues like speaker's tonality to be understood as sarcastic (Caucci and Kreuz, 2012).

Research has shown that tone of voice (Rockwell, 2000; Cheang and Pell, 2008; Woodland and Voyer, 2011), facial micro-expressions (Mishra et al., 2016), and the broader temporal or situational context provide essential information for disambiguating sarcasm. This task is challenging even for humans, as it requires reasoning over cues from diverse sources (Farha et al., 2023; Farabi et al., 2024a). Thus, sarcasm detection becomes a multimodal reasoning challenge which relies on perceiving and integrating subtle cues across modalities.

Previous work (Farabi et al., 2024b) has shown that for multimodal models, performing complex tasks such as sarcasm detection from conversational videos is challenging because unlike traditional multimodal tasks such as captioning or visual question answering, sarcasm detection relies on inferences about underlying intent rather than simple integration of factual information. Thus, interpreting

sarcasm requires pragmatic reasoning in order to resolve the incongruity between the literal meaning of an utterance and non-verbal cues.

Most available datasets provide only coarse-grained binary labels and lack annotations that localize relevant multimodal cues or outline the reasoning process involved, revealing the gap in existing resources for evaluating how multimodal models reason about sarcasm. To bridge this gap, we propose **MUStReason** which provides fine-grained annotations useful for investigating reasoning about sarcasm in Video-Language Models (VideoLMs). This work is the first to introduce a diagnostic benchmark which provides access to reasoning-aligned annotations enabling recognition of key areas where models fail (perceptual issues or faulty reasoning) and provides an in-depth evaluation of VideoLMs' performance in detecting sarcasm. In addition, we propose PragCoT, a pragmatic reasoning framework, which enables derivation of implied intentions from conversational videos, a key step towards detecting sarcasm. Our key contributions are:

- We introduce MUStReason to enable investigation of perceptual and reasoning gaps in VideoLMs while identifying sarcasm.
- We introduce PragCoT, a pragmatic reasoning framework to interpret the literal meaning of the combined modalities as well as underlying intent conveyed by contextual cues.

- We benchmark sarcasm classification in VideoLMs with and without structured reasoning.

2. Related Work

2.1. Multimodal Sarcasm Detection

Sarcasm is a complex rhetorical device where the intended meaning contrasts with the literal surface form. According to Gricean implicature theory (Grice, 1975), sarcasm emerges when a speaker violates conversational norms, prompting listeners to infer alternative intent. For instance, the utterance *"I love the idea!"* may be interpreted as sarcastic when non-verbal cues such as facial expressions and tonality of the speaker suggest the opposite meaning (Attardo, 2000). Inferring sarcasm requires perceiving and reasoning over these non-verbal cues (Farabi et al., 2024a; Caucci and Kreuz, 2012). Studies show that over 76% of ironic utterances in face-to-face settings include explicit verbal, paraverbal, or non-verbal markers (Athanasiadou and Colston, 2020). Facial expressions and gestures help humans to detect sarcasm (Giustolisi and Panzeri, 2021), with indicators such as eye-rolls (Tabacaru and Lemmens, 2014) and mouth movements (Rockwell, 2000). Prosodic features and vocal cues like slower tempo and lower pitch also contribute significantly (Cheang and Pell, 2008).

Early computational approaches focused on textual sarcasm detection in social media (Hazarika et al., 2018; Poria et al., 2016), but later works recognized the importance of non-verbal context. MUSTARD (Castro et al., 2019) introduced a video-based sarcasm dataset, followed by MUSTARD++ (Ray et al., 2022), which incorporated emotion and modality-based sarcasm categories. These datasets enabled models to leverage facial, auditory, and contextual signals, prompting the development of multimodal fusion techniques using attention (Pramanick et al., 2022; Aggarwal et al., 2023), contrastive learning (Zhang et al., 2021), and adaptations of transformer-based architectures (Pan et al., 2020; Babanejad et al., 2020). However, current models often struggle with sarcasm detection for cases that rely primarily on non-verbal cues (Bhosale et al., 2023), partly due to datasets using only binary sarcasm labels without detailed annotations of relevant cues. This hinders error analysis and targeted improvements. To address this gap, we introduce fine-grained multimodal annotations that describes cues relevant for sarcasm and how to reason over them to detect sarcasm.

2.2. Video Understanding with LLMs

VideoLMs encode sampled video frames and align the extracted visual features with corresponding text, which are then processed by large language

models (Sun et al., 2019; Lin et al., 2024; Maaz et al., 2024b,a; Zhang et al., 2023a). Early models like VideoBERT (Sun et al., 2019) were followed by later improvements in alignment using unified visual features in VideoLLaVA (Lin et al., 2024), segment-wise frame sampling (Maaz et al., 2024b), and Q-Former spatial-temporal modeling in VideoLLaMA (Zhang et al., 2023a). Instruction-tuned models like VideoChatGPT (Maaz et al., 2024a) enable conversational video understanding. However, a key limitation of most VideoLMs is the lack of audio processing (Cheng et al., 2024). This restricts their ability to capture audio cues which play an essential role in tasks like sarcasm detection. Recent models like VITA 1.5 (Fu et al., 2025b) and Qwen2.5Omni (Xu et al., 2025b) have introduced audio encoding capabilities. However, VideoLMs continue to struggle with tasks that require higher-level reasoning (Li et al., 2025), as the integration of temporal, visual, and auditory signals remains shallow in many architectures, limiting their ability for nuanced pragmatic interpretation. Our work focuses on enhancing sarcasm detection through pragmatic reasoning and identify gaps that hinder the ability of VideoLMs.

2.3. Multimodal Reasoning

Studies have explored how to elicit reasoning in Large Language Models (LLMs) for complex planning tasks (Huang and Chang, 2023). CoT prompting has emerged as an effective technique to guide models in producing interpretable, stepwise reasoning chains, improving performance on challenging reasoning tasks (Kojima et al., 2022; Wei et al., 2022b). In multimodal settings, early methods used a two-stage pipeline where vision models generated captions that LLMs then reasoned over (Gupta and Kembhavi, 2023; Yang et al., 2023). Recent approaches instead prompt or train models directly on raw modality representations without verbalizing them (Li et al., 2025). For example, Zhang et al. (2023b) combine textual and visual inputs to produce reasoning chains, while Xu et al. (2025a) propose staged reasoning in LLaVA-CoT. Compositional CoT (Mitra et al., 2024) leverages structured representations such as scene graphs to decompose visual scenes before reasoning, further improving interpretability and accuracy. However, reasoning about sarcasm requires considering underlying pragmatics in addition to combining multimodal inputs. In the unimodal setting, LLMs have shown improved sarcasm classification performance when backed by pragmatic reasoning (Lee et al., 2025). Driven by the success of staged reasoning and the necessity of pragmatic insights to interpret sarcasm, we introduce PragCoT, a pragmatic reasoning framework which processes multimodal cues.

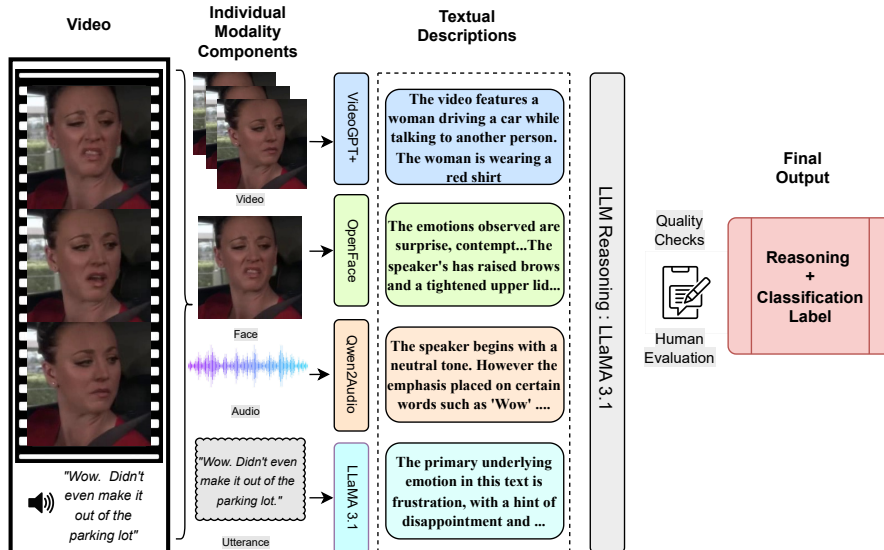


Figure 1: Sarcasm Reasoning Generation Pipeline for MUSTReason

3. MUSTReason

Existing sarcasm detection datasets like MUSTARD (Castro et al., 2019) provide only coarse-grained binary labels and lack fine-grained annotations indicating which specific modalities or cues contribute to making the utterance sarcastic. Without explicit reasoning annotation, it is difficult to diagnose where and why models fail and localize whether the errors stem from wrong perception or from faulty inferential reasoning across the modalities. MUSTReason bridges this gap by providing reasoning-aligned annotations that enable detailed evaluation of modality perception and inference failure which are key factors for assessing and improving pragmatic reasoning in multimodal models.

3.1. Creation

MUSTReason builds on MUSTARD++ Balanced (Bhosale et al., 2023), which includes 691 sarcastic and 674 non-sarcastic video clips. We further add annotations that specify which attributes, individually or in combination, indicate sarcasm in each video. Figure 1 illustrates our automatic process for generating annotations which decomposes sarcasm reasoning into two stages:

Stage 1: Perception of Independent Modalities: We focus on perceiving sarcasm-relevant signals independently from text, audio, and visual modalities. Drawing on insights from linguistic research (Cheang and Pell, 2008), we select a targeted set of cues known to correlate with sarcastic expression, as outlined below.

Statistic	Value
Average Reasoning Length (tokens)	79.51 ± 16.48
Average Unique Tokens	57.16 ± 9.36
Total Unique Tokens	4,346
Total Tokens	107,895
Type-Token Ratio (TTR)	0.0403

Table 1: Descriptive statistics of sarcasm reasoning annotations in MUSTReason.

- **Video:** VideoGPT+ (Maaz et al., 2024b) is used to describe the general scene, setting, and key characters or objects to provide contextual grounding.
- **Face:** We use OpenFace (Baltrusaitis et al., 2018) to identify facial muscle movements and expressions over time that reveal emotional states linked to sarcasm.
- **Audio:** The audio is extracted from the corresponding video using PyDub, which is then passed to Qwen2-Audio (Chu et al., 2024) to capture tonal attributes like mood, pitch changes, and stress patterns that influence the pragmatic meaning.
- **Utterance:** Textual utterances (annotated in the dataset) are passed to LLaMA 3.1 (Dubey et al., 2024) to analyze the emotional content of the spoken words and identify feelings such as mockery, contempt, or frustration.

Stage 2: Pragmatic Reasoning Generation: Once we obtain the individual modality descriptions, we prompt LLaMA 3.1 (Dubey et al., 2024) to collate these, and reason about why the co-occurrence of various attributes or their presence in isolation

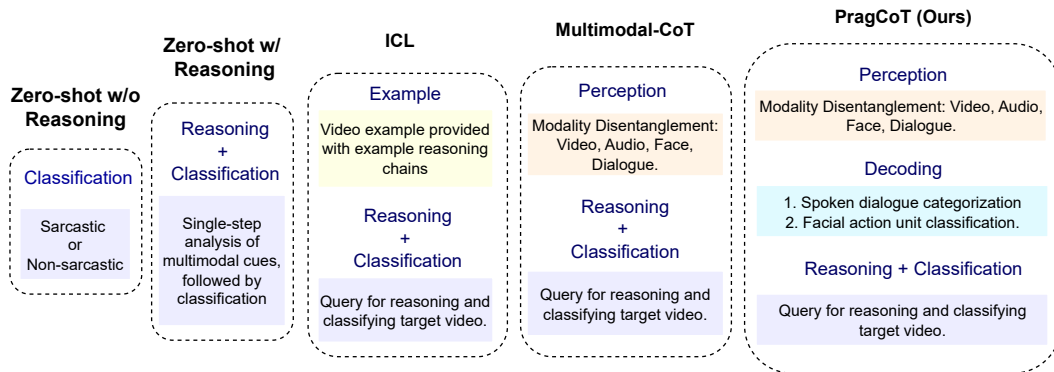


Figure 2: Comparison of prompting strategies for multimodal sarcasm reasoning. PragCoT extends standard CoT by explicitly decoding perceptual cues before reasoning and classification.

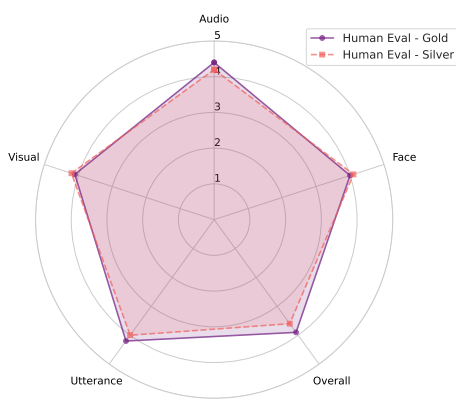


Figure 3: Human Evaluation of Annotation Quality in MUSTReason.

results in the given label being sarcastic or non-sarcastic. The summary contains the most relevant cues which are instrumental in differentiating a sarcastic video from a non-sarcastic one.

Table 1 reports the descriptive statistics of the annotated sarcasm reasoning in MUSTReason. Further details about extraction of modality descriptions, annotation and associated prompts are discussed in Section A.1 of Appendix.

3.2. Evaluation of Generated Reasoning

To evaluate the quality of the automatically generated sarcasm reasoning, we curate a high-quality gold-standard set of 462 instances (approx. 34% of the samples). These samples were checked by two annotators to ensure the correctness of individual modality descriptions, their relevance for sarcasm classification, and the coherence of the overall reasoning. Out of 462 silver instances converted to gold, 56% of the samples required corrections, most of them pertaining to incorrect perception, especially for facial expressions and action units. The

inter-annotator agreement (IAA) was calculated using Edit Distance (0.11) and BERTScore (0.99)¹, which ensured high lexical and semantic alignment between the annotators. In some cases there was a discrepancy between the model’s final prediction and original label, and we excluded these samples.

The remaining 66% of the dataset comprises silver-standard samples (899), which underwent a sanity check to filter out irrelevant content. To estimate how different the silver data is from the gold data, we pass the gold annotations with their corresponding silver version to GPT-4.1 (Fu et al., 2024), prompting it to evaluate the silver reasoning in terms of perceptual correctness and reasoning coherence compared to what is mentioned in the corresponding gold annotation. We find that the silver data achieves a rating of 4.51 out of 5 when compared to the gold data. To further validate the efficacy of silver data, we conduct a user study. We randomly select 10 samples each from the silver and gold sets and ask 26 participants to rate the reasoning on a 1–5 scale. Figure 3 shows the scores achieved by the silver and gold quality reasoning from GPT and human evaluators, supporting the overall quality of MUSTReason

4. Experiments

To benchmark sarcasm detection in VideoLMs, we compare the following strategies with increasing order of guidance as shown in Figure 2:

Zero-Shot without Reasoning: The prompt strictly asks the model to classify the video into either ‘sarcastic’ or ‘non-sarcastic’ class.

Zero-Shot with Reasoning (Han et al., 2023): Given a video, the model is prompted to explain any sarcastic content and assign a label.

In-Context Learning (ICL) (Wei et al., 2022a): A video example along with its corresponding sar-

¹For comparison: Semantic similarity between annotation pairs for different videos are Edit Distance: 0.65, BERTScore: 0.88.

casm reasoning is provided to the model. Given a target video, the model generates a reasoning including mention of the crucial attributes as in the reference reasoning. A practical limitation of this approach is that most VideoLMs are unable to process more than one video in a single conversation.

Multimodal-CoT (Zhang et al., 2023b): A two-stage framework which fuses text and image information to generate a rationale, followed by inferential reasoning. We adapt this to our use case by first extracting individual modality descriptions and then collating them to perform reasoning.

PragCoT (Ours): To enable VideoLMs to reason about sarcasm pragmatically, we propose a structured prompting approach called PragCoT, a variant of Zero-Shot CoT prompting (Kojima et al., 2022). Developing complex reasoning abilities through the integration of information from multiple modalities and the systematic resolution of sub-problems has proven highly effective in multimodal contexts (Zhang et al., 2023b; Fei et al., 2024). However, while multimodal reasoning draws inferences from the combined content of modalities, pragmatic reasoning contains an additional layer of complexity: It relies heavily on contextual cues, social dynamics and the viewer’s interpretation of a situation, focusing more on the underlying intent rather than the literal meaning. PragCoT bridges this gap by introducing a modality-decoding step in addition to the perception and reasoning steps:

1. **Perception**: The model is queried for speaker utterance, acoustic features, generic facial attributes and video backdrop context, separately to disentangle information from the modalities present.
2. **Decoding**: Pragmatic reasoning unlike traditional reasoning aims to derive non-literal meaning conveyed by a situation based on speaker intent and interaction (Grice, 1975). This interpretation heavily relies on how a dialogue is framed (Goffman, 1981) especially in conjunction with facial attributes, the most immediate visual evidence. For instance, when the content of an utterance is not congruent with how it is delivered (e.g., saying "Thank You!" with a deadpan face), the listener’s ability to decode the sarcastic intent is significantly enhanced compared to relying on only words or tone (Attardo et al., 2003). This step, thus, facilitates gathering specific ingredients required for pragmatic reasoning. To capture minute facial attributes, we prompt the model to classify which facial action units from a provided list are present on the speaker’s face. Further, if sarcasm is self-contained in a dialogue, it can render a video sarcastic irrespective of visual or auditory cues. Hence, we

ask the model to categorize dialogues into one of the following classes: neutral, metaphoric, ironic or hyperbolic (Burgers et al., 2016).

3. **Reasoning**: Finally, the model is given detailed instructions about the kind of cues it should attend to, disentangle the incongruity among intents, generate a rationale and aggregate its analysis into a comprehensive reasoning, consisting of the classification label.

For the model choices we use off-the-shelf VideoLMs from the leaderboard (Fu et al., 2025a). Among the 7 models that we evaluate, 5 of them do not support audio. We provide these models explicitly with transcriptions of the speaker utterances, which are available in the original dataset. Details about implementation and prompts for inference using PragCoT are provided in Section A.2 of Appendix.

4.1. Evaluation Metrics

Our evaluation framework assesses performance of VideoLMs with respect to sarcasm classification and quality of generated reasoning. Following previous work (Castro et al., 2019; Zhang et al., 2021), accuracy and macro-F1 are chosen as metrics for binary classification. Next, we leverage the detailed annotations in MUSTReason to assess the perceptual correctness and reasoning coherence of the sarcasm explanations generated by VideoLMs. We conduct an automated GPT4.1 (Achiam et al., 2023) based evaluation to analyze whether the model explanations align with the annotated reasoning in MUSTReason on the basis of correctness of input modality attributes (e.g., spoken words, facial expressions, audio tone), reasoning, and overall understanding of sarcasm.

5. Results and Analysis

MUSTReason serves dual purpose: (1) It acts as a benchmark to help evaluate the reasoning quality of VideoLMs for the task of sarcasm detection. (2) The detailed, modality-focused annotations facilitate the identification of failure cases, reveal whether VideoLMs attend to the relevant attributes, and highlight gaps in perception and reasoning.

5.1. Benchmarking Sarcasm Detection Performance of VideoLMs

Table 2 shows classification results (average of 3 runs) across 5 methods for 7 VideoLMs. PragCoT resulted in higher F1 scores with statistically significant ($p\text{-value}=0.025 < 0.05$) improvements of up to 20% compared to zero-shot classification, about 17% with respect to ICL and 2.4% compared to

Model	Modalities	ZS		ZS		ICL		Multimodal CoT		PragCoT	
		w/o Reasoning		w/ Reasoning		Acc	F1	Acc	F1	Acc	F1
		Acc	F1	Acc	F1						
Video-LLaVA (Lin et al., 2024)	T + V	49.6	33.9	52.2	42.9	-	-	50.3	33.7	50.2	34.1
VideoGPT+ (Maaz et al., 2024b)	T + V	49.7	35.0	50.5	39.8	-	-	50.7	33.6	56.4	36.1
ShareGPT4Video (Chen et al., 2024)	T + V	50.9	46.4	47.8	47.6	-	-	54.9	54.1	55.0	52.8
LLaVA-Next-Video (Zhang et al., 2024)	T + V	50.2	38.5	49.6	35.7	50.0	33.3	52.3	48.4	54.8	53.2
Qwen2.5VL (Bai et al., 2025)	T + V	49.6	39.5	50.6	35.4	47.8	32.3	58.0	57.2	57.4	55.3
VITA (Fu et al., 2025b)	T + V + A	49.6	33.7	49.5	33.6	-	-	49.4	33.2	49.6	33.4
Qwen2.5Omni (Xu et al., 2025b)	T + V + A	49.6	39.5	48.9	34.3	55.0	42.2	57.0	56.9	59.5	59.3

Table 2: Performance of Video-Language Models on Sarcasm Classification using CoT Prompting (Acc: Accuracy, F1: macro F1-Score).

Components	Acc	F1
PragCoT w/o Dialogue Decoding	56.6	56.6
PragCoT w/o Face Decoding	59.0	58.7
PragCoT	59.5	59.3

Table 3: Ablation of the Decoding step in PragCoT for Qwen2.5Omni.

Multimodal-CoT. We observe that VideoLMs perform quite poorly in the zero-shot classification setting. In fact for few models (Qwen2.5VL, LLaVA-Next-Video and Qwen2.5Omni), the F1-score goes down when the model generated an accompanying reasoning compared to zero-shot classification without reasoning. Low performance in the zero-shot setting suggests that the models might not be attending to the appropriate cues during classification. Alternatively, it could indicate the model's dependence on implicit heuristics for sarcasm detection, with a limited ability to articulate intermediate reasoning (Anantha Ramakrishnan et al., 2025). ICL shows slight performance gains over zero-shot classification for the Qwen2.5Omni model only, indicating that providing an example might be helpful but not entirely reliable. Multimodal-CoT, owing to its multi-step reasoning process achieves higher scores than the single-step methods. PragCoT achieves best performance for majority of the models highlighting how crucial the decoding step is for better perception-reasoning alignment. We provide evidence of empirical improvement on addition of the decoding step in Table 3.

Qualitative Evaluation of Reasoning: In addition to assessing sarcasm classification performance, we also evaluate the generated reasoning using MUStReason by comparing the generated reasoning with the annotated data. GPT4.1 judges the sarcasm reasoning based on a set of guidelines, ensuring the credibility of the perceptual cues and the overall relevance of the reasoning. Figure 4 displays the ratings provided by GPT for the overall quality of the reasoning and description of individual modalities within the reasoning generated by the discussed prompting techniques compared to the gold reasoning. We observe a consistent im-

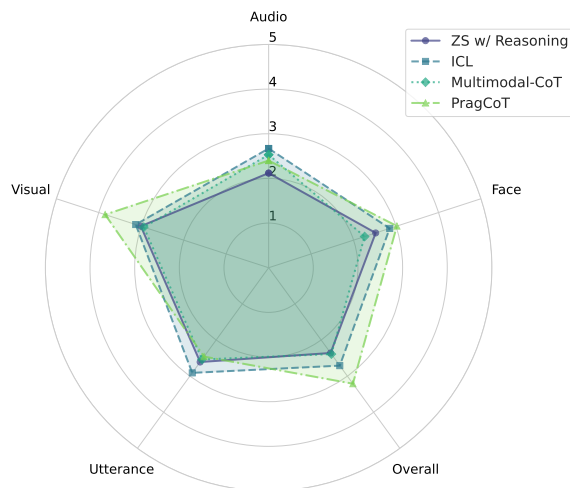


Figure 4: Qualitative Evaluation of Model Generated Reasoning.

provement across most aspects of the reasoning generated via PragCoT. The visual modality descriptions receive the highest ratings, likely because VideoLMs are predominantly trained on scene understanding. In contrast, they show limited ability in capturing intent and audio tone.

Table 4 presents examples of sarcasm reasoning generated by different prompting techniques for the Qwen2.5Omni model. We observe that zero-shot reasoning is prone to hallucinations; although Qwen2.5Omni can interpret audio, its reasoning often claims the absence of any audio input. ICL vividly articulates the factors contributing to sarcasm but tends to include irrelevant factual arguments. This suggests that exposure to examples might teach the model to perceive the essential cues, but it still fails to build meaningful association between them. Multimodal-CoT incorporates standard perceptual cues such as speech content but often overlooks finer indicators like facial expressions, that hinders coherent reasoning as highlighted in


 <p style="text-align: center;">Utterance: "Interesting. You see betrayal in others, but not yourself."</p>
<p>Gold: The video is sarcastic because of the speaker's utterance "you see betrayal in others, but not yourself", which implies hypocrisy and double standards, conveying a negative sentiment with a hint of frustration and anger. The tone is assertive and slightly confrontational, and the facial expressions, such as raised brows and forced smile, supports the sarcastic tone. The irony and criticism in the statement, combined with the speaker's composed and unemotional delivery, also contribute to the sarcastic label.</p>
<p>ZS w/ Reasoning: To analyze the video for sarcasm, let's break down the elements: The setting appears to be a casual dining area, possibly a cafeteria or a similar communal space. There are other people around, some eating and others engaged in conversation. The environment seems relaxed and informal. The audio tone is not explicitly provided, but we can infer from the visual context that the speaker might be using a neutral or slightly casual tone, as there are no indications of an exaggerated or mismatched delivery that would suggest sarcasm. The main character is seen speaking while holding a fork and a drink. His facial expression does not show any obvious signs of sarcasm such as raised eyebrows or a forced smile. Instead, he appears to be engaged in a normal conversation. The dialogue is not provided in the description, so we cannot determine the sentiment directly. However, based on the context, it seems to be a neutral conversation without any indication of sarcasm. Given the lack of any clear indicators of sarcasm in the audio tone, facial expressions, and dialogue sentiment, the scene does not provide strong evidence for sarcasm.</p>
<p>ICL: The video is sarcastic. The speaker's tone is likely to be mocking or dismissive, which is a common characteristic of sarcasm. The dialogue sentiment is negative, as the speaker is pointing out a perceived flaw in the other person's ability to recognize their own betrayal. The facial expressions and AUs, such as raised eyebrows or a forced smile, would likely contradict the spoken words, further indicating sarcasm. The setting appears to be a casual environment, possibly a workplace or social gathering, which could contribute to the context of the sarcastic remark.</p>
<p>Multimodal-CoT: The video segment is likely to be ironic or sarcastic. The serious tone and content of the speech, combined with the incongruity between the serious tone and the casual setting, suggest that the speaker is making a pointed observation about the other person's behavior in a way that is intended to be taken seriously but may also be interpreted as ironic or sarcastic.</p>
<p>PragCoT: The video segment is sarcastic. The utterance cues, particularly the use of "interesting" and the direct contradiction in the statement, clearly indicate sarcasm. The audio cues support this with an accusatory and slightly dismissive tone, as well as a slightly raised pitch at the end of the sentence. The facial expression cues, while neutral to slightly concerned, do not negate the sarcastic tone of the speech. The overall interaction between the tone and the facial expression further reinforces the sarcastic nature of the statement.</p>

Table 4: Example of sarcasm reasoning generated by Qwen2.5Omni model for different prompting strategies. The texts marked in red, blue and olive indicate wrong attribute prediction, hallucination and uncertainty, respectively.

the example. In contrast, PragCoT integrates essential cues, including facial expressions, achieving more accurate sarcasm detection by distinguishing between literal and implied meanings.

5.2. Analyzing Perception and Reasoning in VideoLMs for Determining Sarcasm

We observed that even though classification preceded by reasoning improves accuracy, VideoLMs perform quite poorly in classifying sarcasm compared to state-of-the-art models (Dong et al., 2025) fine-tuned for the task. VideoLMs are composed of encoders which process image frames, speech and text, followed by an LLM decoder capable of reasoning and decision-making. This leads us to examine whether these models fail in accurately perceiving multimodal cues essential for detecting sarcasm or struggles to construct the reasoning chain. MUSTReason lays a ground for investigation by providing gold-standard sarcasm reasoning annotations, which could be qualitatively and quantitatively compared with model-generated responses to identify the loopholes. Upon analyzing the failure cases, we identify two types of errors as observed in Figure 5:

Failure in both perception and reasoning: In Example 1, we observe that the model vaguely assigns a *neutral* tag to all the modalities, without

describing what each modality individually conveys. Lack of concrete details results in minimal reasoning by the model. Comparing with the gold annotation it is evident how the mocking utterance, in conjunction with the speaker's expressions and actions, supports the conclusion that the video is sarcastic. This highlights the importance of accurate perception as a prerequisite for reasoning when tackling complex tasks like sarcasm detection. Thus, if the model fails to attend to key perceptual cues, the reasoning thread gets disrupted, resulting in an incorrect prediction.

Failure in reasoning with correct perception: For Example 2, the model correctly identifies *expression of frustration* and *furrowed eyebrows* of the speaker. The gold annotation shows how reasoning is grounded in these attributes. In contrast, a preceding spurious cue generated by the model subsequently leads to an under-confident reasoning chain. It completely overlooks the correct cues it had identified earlier, leading to an incorrect prediction. In this type of error, even though the model correctly perceives information from the different modalities, it lacks confidence in the reasoning process or fails to infer the association between the cues that would lead to the correct label.

We further leverage the MUSTReason dataset to evaluate the model's reasoning quantitatively. Firstly, we compare the model's response with the gold reasoning using semantic metrics like



Figure 5: Qualitative and quantitative analysis of model-generated reasoning using annotations from the MUSTReason dataset. These annotations help in identifying whether the model fails in perception, reasoning or both when identifying sarcasm. **Bold** words in Gold indicate the attributes expected to be present in Predicted. Text marked in **green** and **red** indicate **correct** and **wrong** predictions respectively. Tags <P> and <R> represent *perception* and *reasoning* respectively.

BERTScore (Zhang et al., 2019), METEOR (Banerjee and Lavie, 2005) and RougeL (Lin, 2004) and observe that the inclusion and exclusion of cues and reasoning steps reflects in the scores. Since reasoning might have low semantic similarity while still conveying the same meaning, we use GPT as a judge to compare the model generated response with gold annotation. The evaluation rates the presence of individual modality descriptions and reasoning in a scale of 1-5, from highly dissimilar to highly similar. For Example 1, which falls short in both perception and reasoning, the scores are 1, further establishing the reason of failure. In Example 2, while there is some similarity in video context and facial expressions, the overall rating remains low due to incorrect construction of reasoning chain. To examine how reasoning quality degrades from correct to incorrectly classified samples, we calculate the mean of the ratings provided by GPT for overall reasoning quality for each class, obtaining scores of 3.02 for correct and 1.64 for incorrect predictions, on a 5-point scale.

Table 8 and 9 in Section A.3 of Appendix show the variation in perceptual ratings and reasoning

Model	Acc	F1
<i>Large Language Models (LLMs)</i>		
Llama-3.1-8B-Instruct (Dubey et al., 2024)	59.0	57.6
Qwen3LM-8B (Yang et al., 2025)	60.5	60.0
Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)	61.8	61.7
GPT-5	65.6	65.6
<i>Video-Language Models (VideoLMs)</i>		
LLaVA-Next-Video	60.6	60.4
Qwen2.5VL	61.8	61.7
Qwen2.5Omni	60.7	59.7

Table 5: Sarcasm classification performance of LLMs vs VideoLMs when provided with quasi-perfect perception.

scores for correct and incorrect model predictions for all 1365 samples. The prompt for GPT-based evaluation has also been included.

Do LLMs outperform VideoLMs in sarcasm detection when provided with quasi-perfect perceptual cues?

During error analysis we observe that incorrect perception misguides the model to build an incorrect reasoning chain. This raises the question: given near-perfect modality descriptions derived

individually from *modality* → *text* models, how well can the VideoLMs reason? To investigate this, we undertake an approach where the model is supplied with individual modality descriptions and tasked to only reason and explain. This is typically employed for LLMs that lack the ability to process audio or visual inputs. In our setup, we extend this method to VideoLMs as well. This helps in assessing the reasoning ability of these models. Table 5 shows a marginal increase in classification scores for Qwen2.5Omni compared to Table 2 indicating the model suffers from a reasoning bottleneck even when perception is improved. For LLaVA-Next-Video and Qwen2.5VL, the increase is higher, suggesting a substantial gap in perception as well. This proves that finer-grained perception is an essential first step to reason about complex pragmatic phenomenon like sarcasm. LLMs achieve performance comparable to VideoLMs, with GPT-5 being the best reasoner, beating the open-source LLMs. However, the overall classification performance remains low for both LLMs and VideoLMs, indicating that these models are still far from demonstrating pragmatic reasoning capabilities.

6. Conclusion

In his work, we analyze the performance gaps of existing open-source VideoLMs in perceiving, reasoning and detecting sarcasm, which requires interpreting and disambiguating conflicting information gathered from diverse modalities, we introduce the MUSStReason diagnostic benchmark which contains samples annotated with sarcasm reasoning. Furthermore, we devise a multi-stage structured prompting approach, PragCoT, in order to incorporate pragmatic reasoning abilities in VideoLMs, which is key to understanding and inferring underlying sarcastic implications in videos. While PragCoT provides one example of how MUSStReason can be used to evaluate reasoning behavior, the annotations can support evaluation of any generic reasoning framework and similar reasoning annotations can be generated to assess tasks like humor detection and figurative language understanding.

7. Limitations

A primary limitation of our work lies in the sarcasm reasoning annotations being derived from the MUSStARD++ Balanced dataset. Thus, the contextual cues and non-verbal markers present in the annotations are dataset-specific, preventing generalizability across dyadic conversations which might contain additional cue variations not represented in these annotations. Besides, the dataset is in English. Since sarcasm is influenced by cultural and social dynamics, the monolingual nature of the dataset

misses language-specific sarcasm indicators. In addition, while PragCoT lays the foundation for pragmatic reasoning for sarcasm interpretation by capturing contradictions and incongruities between modalities, integrating ensemble approaches or reinforcement learning based techniques into this framework might help these models generate better reasoning chains. Currently the diagnostic benchmark is limited to sarcasm. Future work would focus on extending it to other tasks like irony and figurative language understanding to make pragmatic reasoning in multimodal models interpretable.

8. Acknowledgement

Varsha Suresh is funded by Deutsche Forschungsgemeinschaft, Funder Id: <http://dx.doi.org/10.13039/501100001659>, SFB 1102: "Information Density and Linguistic Encoding". We sincerely thank Zaynab Batool Reza for her help towards annotating the MUSStReason dataset.

9. References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Sajal Aggarwal, Ananya Pandey, and Dinesh Kumar Vishwakarma. 2023. Multimodal sarcasm recognition by fusing textual, visual and acoustic content via multi-headed attention for video dataset. In *2023 world conference on communication & computing (WCONF)*, pages 1–5. IEEE.
- Aashish Anantha Ramakrishnan, Aadarsh Anantha Ramakrishnan, and Dongwon Lee. 2025. [COR-DIAL: Can multimodal large language models effectively understand coherence relationships?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21277–21297, Vienna, Austria. Association for Computational Linguistics.
- Angeliki Athanasiadou and Herbert L Colston. 2020. *The diversity of irony*, volume 65. Walter de Gruyter GmbH & Co KG.
- Salvatore Attardo. 2000. Irony as relevant inappropriateness. *Journal of pragmatics*, 32(6):793–826.
- Salvatore Attardo, Jodi Eisterhold, Jen Hay, and Isabella Poggi. 2003. [Multimodal markers of](#)

- irony and sarcasm. *Humor-international Journal of Humor Research - HUMOR*, 16:243–260.
- Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th international conference on computational linguistics*, pages 225–243.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. *Openface 2.0: Facial behavior analysis toolkit*. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, page 59–66. IEEE Press.
- 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.
- Swapnil Bhosale, Abhra Chaudhuri, Alex Lee Robert Williams, Divyank Tiwari, Anjan Dutta, Xiatian Zhu, Pushpak Bhattacharyya, and Diptesh Kanojia. 2023. *Sarcasm in sight and sound: Benchmarking and expansion to improve multimodal sarcasm detection*.
- Christian Burgers, Elly A Konijn, and Gerard J Steen. 2016. Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony. *Communication theory*, 26(4):410–430.
- 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.
- Gina M Caucci and Roger J Kreuz. 2012. Social and paralinguistic cues to sarcasm. *Humor*, 25(1):1–22.
- Henry S Cheang and Marc D Pell. 2008. The sound of sarcasm. *Speech communication*, 50(5):366–381.
- Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495.
- Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. *Qwen2-audio technical report*.
- Zhongren Dong, Donghao Wang, Ciqiang Chen, Dong-Yan Huang, and Zixing Zhang. 2025. *Mhsdb: A comprehensive benchmark for multimodal humor and sarcasm detection leveraging foundation models*. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *CoRR*.
- Paul Ekman and Wallace V. Friesen. 1978. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.
- Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024a. A survey of multimodal sarcasm detection. In *IJCAI*.
- Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024b. *A survey of multimodal sarcasm detection*. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24*.
- Ibrahim Abu Farha, Steven Wilson, Silviu Vlad Oprea, and Walid Magdy. 2023. Sarcasm detection is way too easy! an empirical comparison of human and machine sarcasm detection. In *The 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5284–5295. Association for Computational Linguistics.
- Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu.

2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.
- Chaoyou Fu, Haojia Lin, Xiong Wang, Yifan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. 2025b. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *CoRR*.
- Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.
- Beatrice Giustolisi and Francesca Panzeri. 2021. The role of visual cues in detecting irony. In *Proceedings of Sinn und Bedeutung*, volume 25, pages 292–306.
- Erving Goffman. 1981. *Forms of talk*. University of Pennsylvania Press.
- Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.
- Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962.
- Chi Han, Hengzhi Pei, Xinya Du, and Heng Ji. 2023. Zero-shot classification by logical reasoning on natural language explanations. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. [CASCADE: Contextual sarcasm detection in online discussion forums](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. 2025. [Pragmatic metacognitive prompting improves LLM performance on sarcasm detection](#). In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 63–70, Online. Association for Computational Linguistics.
- Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.
- Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024a. [Video-ChatGPT: Towards detailed video understanding via large vision and language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024b. [Videogpt+: Integrating image and video encoders for enhanced video understanding](#).

- Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. [Predicting readers' sarcasm understanding by modeling gaze behavior](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).
- Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.
- Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. [Modeling intra and intermodality incongruity for multi-modal sarcasm detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.
- Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. [A deeper look into sarcastic tweets using deep convolutional neural networks](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1601–1612, Osaka, Japan. The COLING 2016 Organizing Committee.
- Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3930–3940.
- 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*.
- Patricia Rockwell. 2000. Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic research*, 29(5):483–495.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.
- Sabina Tabacaru and Maarten Lemmens. 2014. [Raised eyebrows as gestural triggers in humour: The case of sarcasm and hyper-understanding](#). *The European Journal of Humour Research*, 2(2):11–31.
- 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 *Proceedings of the 29th ACM International Conference on Multimedia*, page 3927–3935.
- Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Jennifer Woodland and Daniel Voyer. 2011. Context and intonation in the perception of sarcasm. *Metaphor and Symbol*, 26(3):227–239.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025a. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.
- 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. 2025b. [Qwen2.5-omni technical report](#).
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.
- Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 543–553.

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*.

Xiaoqiang Zhang, Ying Chen, and Guangyuan Li. 2021. Multi-modal sarcasm detection based on contrastive attention mechanism. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 822–833. Springer.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. [Llava-next: A strong zero-shot video understanding model](#).

Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. 2023b. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

A. Appendix

A.1. Dataset Creation and Annotation

A.1.1. Extracting Individual Modality Descriptions

We employ *modality* → *text* expert models to derive insights from each modality:

1. **Video:** By feeding the videos to VideoGPT+ (Maaz et al., 2024b), we obtain description of the general visual scene, setting, characters and objects present across entire length of the video.
2. **Face:** We use OpenFace (Baltrusaitis et al., 2018) toolkit to extract Facial Action Units (FAU) and derive the expression it corresponds to from the facial muscle movements detected. A sequential heuristic is followed, details of which can be found in Section A.1.2.
3. **Audio:** The audio is extracted from the corresponding video using PyDub, which is then passed to Qwen2-Audio (Chu et al., 2024) for obtaining tonal attributes.
4. **Utterance:** We pass the textual utterances (annotated in the original dataset) to LLaMA 3.1 (Dubey et al., 2024) and obtain insights conveyed by the spoken words.

Following are the prompts used for extracting individual modality descriptions and combining these descriptions along with the gold label to

generate sarcasm reasoning using LLaMA3.1.

Video

SYSTEM: You are an expert at describing videos.

USER: Provide a description of the video that includes information about the actors, their actions, and the scene they are in.

Audio

SYSTEM: You are an audio analysis expert.

USER: Provide a comprehensive analysis of the speaker's vocal dynamics in the following audio. For each segment or the entirety of the audio, comment on:

- **Tone:** (e.g., warm, cold, serious, playful, authoritative)
- **Intonation:** (e.g., melodic, monotonous, emphatic rises, questioning falls)
- **Sentiment:** (e.g., positive, negative, neutral, sarcastic, uncertain)
- **Emotions:** (e.g., happiness, frustration, calmness, anxiety - with evidence from vocal cues)
- **Shifts:** (Note any transitions in tone, intonation, or amplitude, and try to interpret their significance)
- **Amplitude:** (e.g., loud, soft, varying - and how it contributes to the overall delivery)

Below is an example of the style in which you should answer about the provided audio clip.

Example audio utterance: I'm really happy to be here.

Example explanation: The intonation in this part features a slight rising contour on "really happy" and a gentle falling contour on "to be here," which typically conveys enthusiasm and sincerity. The overall sentiment at the start is clearly positive, suggesting joy and contentment. The emotion expressed seems to be happiness or excitement, supported by the bright tone and engaging intonation.

Now comment about the provided audio clip. Be as detailed and specific as possible in your observations, providing justifications based on the auditory cues.

Utterance

SYSTEM: You are a text analysis expert.

USER: Analyze the following text and identify the primary underlying emotion and the overall sentiment expressed. Mention if you sense a hint of sarcasm or not in the text. Do not say "I cannot provide an analysis of a text". Do not start with "Based on the provided text". Do not say that the text lacks context or anything similar. Provide an analysis always and whatever you can. {utterance text}

Pragmatic Reasoning Generation

SYSTEM: You are a summarization expert.

USER: You have been given a textual description of a video. The description contains the speaker utterance, the underlying emotion or sentiment conveyed by the speaker utterance, the video context description, the speaker's tonality description, the speaker's facial description and possible emotions. The description has also been labeled as "sarcastic" or "non-sarcastic." Your task is to provide a concise yet comprehensive summary, reasoning about how the entire given description can be used to conclude the given label. Include only those parts of the whole description in the summary that are useful to conclude whether the video is sarcastic or non-sarcastic. Do not start your answer with "Based on the provided description" or a similar prefix. {utterance text, utterance sentiment, audio tone description, video context description, speaker facial expressions}

ASSISTANT: The video is <sarcastic/non-sarcastic> because

A.1.2. Derivation of Facial Emotion and Attribute Descriptions for constructing MUSTReason

To extract fine-grained features from the speaker's face, we use the OpenFace (Baltrusaitis et al., 2018) toolkit, which is state-of-the-art for facial behavior analysis. Facial Action Coding System (FACS) (Ekman and Friesen, 1978) is a systematic way to quantize facial muscle movements. Each Action Unit (AU) represents one or more muscle movements and various combinations of these AUs can represent an expression or emotion. Since, a

video has multiple frames and each frame can have non-speaker characters, we take a sequential approach to detect AUs and facial emotions:

1. For each frame of a video, we first detect and crop the speaker faces using TalkNet (Tao et al., 2021). To improve resolution for precise AU detection, the cropped frames are superpixelized through Real-ESRGAN (Wang et al.).
2. Next, we obtain the AU activation values from OpenFace for each of the frames. Since, a speaker's facial expression is prominent only for a certain subset of the frames, we filter out the redundant frames having AU activation values below a certain threshold (< 0.5). This gives a set of AUs having the highest activation values across the frames.
3. The set of AUs are then replaced by their corresponding description of the muscle movements (Table 6), which are then aggregated into a single description through the template "*The speaker's face shows ...<list of AU descriptions>*".
4. We refer to Table 7 to assign expression labels to the corresponding unique combination of AUs obtained in step 2. Finally, they are plugged into a similar template "*The possible emotions observed along different segments of the video are: ...<list of emotions>*".

Facial Action Unit	Description
AU01	Inner Brow Raised
AU02	Outer Brow Raised
AU04	Brow Lowered
AU05	Upper Lid Raised
AU06	Cheek Raised
AU07	Lid Tightened
AU09	Nose Wrinkled
AU10	Upper Lid Raised
AU12	Lip Corner Pulled
AU14	Dimple
AU15	Lower Lip Depressed
AU17	Chin Raised
AU20	Lip Stretched
AU23	Lip Tightened
AU25	Lips Parted
AU26	Jaw Dropped
AU45	Blink

Table 6: Facial Attribute Description for Facial Action Units.

A.1.3. Data Filtering and Annotation

The sarcasm explanations were generated automatically by individual modality expert models and

Facial Expression	Facial Action Units
Happiness	AU06, AU12, AU14
Worry	AU28, AU20
Surprise	AU01, AU02, AU05, AU26
Sadness	AU04, AU01, AU14, AU15
Doubt	AU25
Contempt	AU12, AU10, AU15, AU17
Anger	AU04, AU05, AU07, AU23, AU10, AU17
Fear	AU01, AU02, AU04, AU05, AU07, AU20, AU26

Table 7: Mapping between Facial Expressions and Facial Action Units

collated using LLaMA 3.1. However, models are often known to hallucinate and any misinformation generated at the individual level, might travel down to the sarcasm summary. Since human annotation is very time-consuming and expensive, we curate two types of dataset samples - *silver-standard* and *gold-standard* data. The samples in the silver-standard data go through a basic sanity check to ensure the model has not generated gibberish or irrelevant content. On the other hand, each of the sample in the gold-standard data are human-annotated to ensure correctness of the mentioned multi-modal attributes, their overall relevance in sarcasm classification and quality of the generated explanation. The high quality gold data ensures that model generated explanation quality would be evaluated on a standard benchmark. The explanations were also checked for spelling and grammatical errors. For few instances, we observed that LLaMA 3.1 generated contradictory explanations. For example, when provided with the ground truth label *sarcastic* for reasoning, the model indicated the instance was *non-sarcastic* and justified its response accordingly. To ensure the accuracy of these cases, we manually reviewed the corresponding samples to identify any inconsistencies or explanation errors. Upon inspection, we found that in four cases, LLaMA's predictions were actually correct and the ground truth labels were incorrect. These samples were excluded. The resulting dataset consists of 462 gold and 899 silver samples, each set having a nearly balanced distribution of sarcastic and non-sarcastic classes. The MUSTARD and MUSTARD ++ Balanced dataset provides the raw videos and sarcasm annotations. Each video clip has a unique key. The dataset we provide contains sarcasm reasoning annotated against these keys.

A.1.4. Data Constraint for Sarcasm Detection

High-quality videos are extremely difficult to source. Most publicly released video datasets only offer video frames instead of the raw videos due to ethical and copyright issues. Perception-grounded

tasks like sarcasm detection relies on scene continuity to spot dynamic changes with time. As discussed before, audio also plays a vital role in explaining sarcasm. Besides, sarcasm is a very distinct genre which is hard to find in commonly available videos on YouTube. A bulk of sarcastic clips can only be fetched from a limited number of TV shows, especially sitcoms. Quite often, these videos are of extremely low resolution. Cropping faces from such videos causes blurring and increases the difficulty of extracting facial information. Identifying and annotating clips containing sarcastic content is time-consuming, labour-intensive and expensive. Besides, there is always some risk of bias and cultural perspective involved as to how humans interpret sarcastic intent. For instance while annotating the MUSTARD ++ Balanced dataset for sarcasm explanation, we in fact found a few samples having wrong class labels.

A.2. Experiments

A.2.1. Implementation Details

For our experiments, we utilize the 7B parameter variant of every VideoLM, except for ShareGPT4, which offers a 8B variant. We conducted inference on NVIDIA Tesla A40 (48GB) and A100 (80B) GPUs, depending on the memory footprint of each model. The max number of new tokens to be generated was set to 512 for every model. For the ICL experiment, except for the models whose classification scores are included in the table, all the other models do not accept more than one video input in a single turn of the conversation. Hence, the scores for these models are empty.

A.2.2. Prompts for Inference using PragCoT

Perception: Video

USER: Summarize what is happening visually in the scene.

Perception: Audio

USER: Describe the speaker's utterance tone, pitch, rhythm, and prosodic cues in the audio/speech.

Perception: Utterance

USER: What does the speaker say?

Perception: Face

USER:

1. Describe the speaker's facial expression and facial emotion in details.

2. Choose one or more attributes from the following list which can be observed on the speaker's face:

['Inner brow raised','Outer brow raised','Brow lowered','Upper lid raised','Cheek raised','Lid tightened','Nose wrinkled','Upper lid raised','Lip corner pulled','Dimple','Lower lip depressed','Chin raised','Lip stretched','Lip tightened','Lip parted','Jaw dropped','Blink']

Decoding: Facial Action Unit Classification

USER: Choose one or more attributes from the following list which can be observed on the speaker's face:

1. Inner brow raised
2. Outer brow raised
3. Brow lowered
4. Upper lid raised
5. Cheek raised
6. Lid tightened
7. Nose wrinkled
8. Upper lid raised
9. Lip corner pulled
10. Dimple
11. Lower lip depressed
12. Chin raised
13. Lip stretched
14. Lip tightened
15. Lip parted
16. Jaw dropped
17. Blink

Decoding: Utterance Classification

USER: Interpret the dialogue to understand whether it is neutral, metaphorical, ironic or hyperbole : DIALOGUE: <utterance>

Reasoning

SYSTEM: You are a highly intelligent and helpful assistant who can understand text, audio and video. Your task is to answer questions thoughtfully. Always try to explain your reasoning process step-by-step before giving a final answer, especially for complex queries. Be concise when asked for a summary, but elaborate when more detail

is requested.

USER:

TASK: Explain in detail why the given video segment is sarcastic or non-sarcastic.

REQUIREMENTS:

- Go step-by-step for **each cue**, one modality at a time.
- **Explicitly mention and use all modality cue descriptions** listed above in your explanation.
- Compare and contrast cues across modalities.
- Consider interactions (e.g., over-enthusiastic tone with flat expression).
- Do not omit any modality or cue.

A.3. Quantitative Evaluation of Model Generated Reasoning

Table 8 and 9 show the variation in perceptual ratings and reasoning scores for correct and incorrect model predictions for all 1365 samples.

Predictions	Utterance	Video	Audio	Face
Incorrect	1.41 \pm 0.79	2.56 \pm 0.89	1.89 \pm 0.83	1.80 \pm 0.79
Correct	3.52 \pm 0.93	3.37 \pm 0.88	2.99 \pm 0.98	2.43 \pm 0.83

Table 8: GPT ratings (scale: 1-5) for perceptual correctness for correct and incorrect predictions of Qwen2.5Omni model for all samples.

Predictions	BERTScore	METEOR	RougeL	GPT
Incorrect	0.87 \pm 0.07	0.24 \pm 0.06	0.24 \pm 0.05	1.64 \pm 0.73
Correct	0.88 \pm 0.01	0.20 \pm 0.06	0.20 \pm 0.04	3.02 \pm 0.86

Table 9: Reasoning quality for correct and incorrect predictions of Qwen2.5Omni model for all samples. Metric scales: GPT: 1-5, Others: 0-1.

Prompt for GPT-based Evaluation

Compare a "GENERATED" explanation to a "GOLD" explanation for sarcasm based on these components:

- (1) Description of video context (setting, environment, etc.)
- (2) Facial expressions and action units
- (3) Audio tone of the speaker
- (4) Intent of speaker utterance.

Carefully analyze each component, rate their similarity from 1 to 5 (where 1 = totally dissimilar, 2 = somewhat dissimilar, 3 = mixed similarity, 4 = somewhat similar, 5 = totally similar), then provide an overall similarity rating focused specifically on reasoning about sarcasm. You must explain your reasoning for each component before giving the rating. The overall rating should be based on your analysis of the three components and reflect the quality of the silver explanation's reasoning compared to the gold. Be thorough in your analysis before outputting any ratings.

Output your answer as a JSON object, structured with "reasoning" preceding all ratings, and the final overall rating at the end. For each reasoning field, provide a short paragraph (2-4 sentences) explaining your comparison. For each rating field, use only an integer from 1 to 5.

Example output format:

"Context_Reasoning": Analysis comparing video context in generated vs. gold; state if details, accuracy/relevance, and completeness about setting environment are similar or different, and why.,

"Context_Rating": 1-5,

"Facial_Reasoning": Analysis comparing facial expressions/action units in generated vs. gold; state if key actions, detail level, and interpretive accuracy are similar or different, and why.,

"Facial_Rating": 1-5,

"Audio_Reasoning": Analysis comparing audio tone in generated vs. gold; state if descriptors, nuance, or identification are similar or not, and why.,

"Audio_Rating": 1-5,

"Intent_Reasoning": Analysis comparing utterance intent in generated vs. gold; state if descriptors, nuance, or identification are similar or not, and why.,

"Intent_Rating": 1-5,

"Overall_Reasoning": Summarize how the silver explanation's reasoning about sarcasm, across all three components, compares to the gold. Explain which aspects are most/least similar and the main strengths/weaknesses.,

"Overall_Rating": 1-5

(Reminder: Output MUST be in JSON as above, with reasoning before each rating and only integer ratings 1-5.)