Scale Is All You Need 😕: Analyzing Modality Interaction and Speaker Intent Without Fine-Tuning

Animesh Gurjar and Nikhil Krishnaswamy
Situated Grounding and Natural Language (SIGNAL) Lab
Department of Computer Science
Colorado State University
Fort Collins, CO, USA

Abstract

Understanding sarcasm requires integrating cues from language, voice, and facial expression. Recent work has achieved impressive results using large multimodal Transformers, but such models are computationally expensive and often obscure how each modality contributes to the final prediction. This paper introduces a lightweight, interpretable framework for multimodal sarcasm detection that combines frozen text, audio, and visual embeddings from pretrained encoders through compact fusion heads. Using the MUStARD++ Balanced dataset, we show that early fusion of textual and acoustic features improves over the best unimodal baseline. Character-specific evaluation further shows that sarcasm expressed through overt prosodic and visual cues is substantially easier to detect than monotone, context-dependent sarcasm. Additionally, we evaluate generalization to different characters through leave-one-speaker-out (LOSO) experiments and run ablation-style transfer experiments on two speakers with similar sarcasm distributions. These findings demonstrate that effective multimodal sarcasm understanding can emerge from frozen, resource-efficient representations without large-scale fine-tuning, emphasizing the importance of modality interaction and delivery style rather than model scale.

1 Introduction

Sarcasm is a complex communicative phenomenon in which speakers express meanings that differ from, or even contradict, the literal interpretation of their words. Accurately detecting sarcasm remains difficult for computational models because it depends on subtle interactions among lexical content, tone, and facial expression. While most humans can effortlessly interpret sarcastic intent by integrating these cues, computational systems often fail when sarcasm departs from explicit linguistic markers and relies instead on delivery or shared background knowledge.

Recent advances in large multimodal Transformers have achieved strong benchmark results on sarcasm and humor detection, but these models are resource-intensive and opaque. They require significant fine-tuning, large GPU memory, and domain-specific supervision, which limits reproducibility and interpretability. Moreover, their performance gains often stem from model scale rather than improved understanding of *how* sarcasm manifests across modalities (Bhosale et al., 2023; Zhang et al., 2024; Dong et al., 2025). As a result, it remains unclear whether more compact architectures can achieve comparable performance on sarcasm detection while providing greater insight into the multimodal nature of sarcasm.

An additional challenge often overlooked in prior work is speaker variability: sarcasm is not expressed uniformly across individuals, and personal delivery style and intent can substantially affect how multimodal cues signal sarcasm.

In this work, we investigate whether reliable sarcasm recognition can emerge from *lightweight, interpretable architectures* that use frozen pretrained encoders instead of end-to-end fine-tuning. Our approach isolates the contribution of each modality—text, audio, and visual—by fusing compact representations from RoBERTa, wav2vec 2.0, and OpenFace through shallow classifiers such as logistic regression and small MLPs. By freezing the encoders, we remove the confounding influence of representational drift during fine-tuning, ensuring that observed differences arise purely from modality interaction and fusion design. This design also enables controlled speaker-level analysis, allowing us to examine how the same multimodal representations behave across different characters with distinct sarcasm styles. In summary, this paper makes the following contributions:

- We present a resource-efficient multimodal sarcasm detection framework using pretrained encoders and compact fusion heads.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 483–492

- We systematically evaluate unimodal and multimodal configurations under five-fold grouped cross-validation on the MUStARD++ Balanced dataset.
- We introduce a speaker-specific comparison of delivery styles, revealing that expressive, intentional sarcasm is easier to recognize than monotone, context-bound sarcasm.
- We evaluate cross-speaker generalization, using leave-one-speaker-out (LOSO) evaluations and targeted ablation-style cross-speaker experiments where certain speaker-specific data is withheld during training.

These contributions provide a transparent baseline for multimodal sarcasm detection and demonstrate that meaningful multimodal understanding can emerge without large-scale task-specific fine-tuning. Our findings highlight that effective sarcasm recognition is not solely a function of model scale, and can also be facilitated by a better understanding of how modalities interact during sarcasm delivery, at less overall computational cost. Our preprocessing pipeline, including generated WIDE features and speaker splits, is available at `https://github.com/csu-signal/multimodal_sarcasm_detection`.

2 Related Work

2.1 Text-Based vs. Multimodal Sarcasm Detection

Early sarcasm detection relied primarily on textual cues such as sentiment polarity shifts, lexical incongruity, and pragmatic markers (Joshi et al., 2017). Pretrained transformers like BERT and RoBERTa substantially improved performance by modeling contextual semantics and discourse-level dependencies (Zhou et al., 2024). Recent studies have also examined zero- and few-shot prompting with large language models (LLMs), demonstrating strong reasoning over text but limited grounding in paralinguistic or visual cues (Zhang et al., 2024), motivating multimodal integration.

Multimodal Datasets and Benchmarks The MUStARD dataset (Castro et al., 2019) established the first multimodal benchmark for sarcasm recognition, aligning textual, acoustic, and visual features from television dialogue. MUStARD++ (Ray et al., 2022), expanded the corpus with balanced sarcasm labels, richer speaker metadata, and improved cross-modal synchronization.

Beyond MUStARD and its variants, several large-scale resources provide broader benchmarks for multimodal irony and humor. MHSDB (Dong et al., 2025) integrates multilingual sarcasm and humor datasets and systematically compares frozen versus fine-tuned encoders, concluding that multimodal combinations yield the most stable cross-domain generalization. SarcasmBench (Zhang et al., 2024) focuses on evaluating LLMs via prompt-based protocols, finding that even state-of-the-art models such as GPT-4 fail to account for audiovisual incongruity, underscoring the need for grounded multimodal reasoning.

Fusion and Incongruity Modeling A core challenge in multimodal sarcasm detection is capturing the incongruity between what is said, how it is said, and how it appears. Raghuvanshi et al. (2025) proposed an intra-modal relation and emotional-incongruity learning network that uses Graph Attention Networks (GATs) to link emotion subspaces within frozen language (BERT), audio (wav2vec 2.0), and visual (ResNet) encoders. This efficiency-driven philosophy aligns closely with our approach. Wu and Zang (2024) introduced the Multi-Scale Adaptive Fusion with Self-Distillation Model (MSAF-SDM), which dynamically reweights modalities and time scales, showing that performance gains stem more from effective fusion than from model scale.

Acoustic and Visual Cues Audio features, such as prosodic variation in pitch, rhythm, and intensity often signals ironic tone even when text is ambiguous, making them important for sarcasm recognition Jose (2025) demonstrated this through a parameter-reduced depthwise CNN that achieves competitive accuracy using only speech features. In contrast, visual features such as facial Action Units (AUs; Ekman and Friesen (1978)), gaze, and head pose, e.g., from OpenFace 2.0 (Baltrusaitis et al., 2018), tend to be noisier in television dialogue data, though they can enhance robustness when combined with audio and text (Castro et al., 2019; Jang and Frassinelli, 2024).

2.2 Lightweight and Efficient Architectures

Most state-of-the-art multimodal sarcasm detectors rely on heavy Transformer backbones with cross-attention, which obscure interpretability and demand significant computational resources. However, recent efforts have instead explored efficiency and modularity. The Hybrid Quantum–Classical Neural Network (HQNN; Phukan et al. (2024))

Model / Method	Dataset	Fusion Type	F1	Params (M)
Raghuvanshi et al. (2025)	MUStARD++	Graph Attention (frozen)	0.749	∼125
Wu and Zang (2024)	MUStARD++	Multi-Scale Adaptive Fusion (fine-tuned)	0.877	∼160
Phukan et al. (2024)	MUStARD++	Quantum–Classical Hybrid	0.712	∼70
Dong et al. (2025)	MUStARD++	CLIP + HuBERT (frozen)	0.774	∼190
Jang and Frassinelli (2024)	MUStARD++	Fine-tuned Transformer	0.630	110
Bhosale et al. (2023)	MUStARD++ Balanced	Early (concat + MLP)	0.736	∼370
Dong et al. (2025)	MUStARD++ Balanced	Utterance (LMF)	0.763	∼1179

Table 1: Recent reported multimodal sarcasm detection system performance on **MUStARD++** or **MUStARD++ Balanced**. Our lightweight model achieves competitive performance with under 1M trainable parameters, compared to 70–190M in prior SOTA systems.

merges quantum circuits with classical deep learning to perform sarcasm, emotion, and sentiment analysis in a compact joint model. Similarly, MHSDB (Dong et al., 2025) show that frozen encoders retain strong transferability when coupled with small fusion heads such as logistic regression or shallow MLPs. Our work builds on this by using entirely frozen encoders and focusing on how modality interaction, rather than parameter count, governs performance.

2.3 Speaker Intent and Personality Effects

Sarcasm is not a uniform phenomenon: its detectability depends strongly on speaker style and intent. While prior datasets include speaker metadata, few studies have examined how delivery differences, such as deadpan versus performative sarcasm, affect model behavior. Even fewer works examine whether sarcasm detectors trained on one set of speakers can generalize to unseen speakers with distinct delivery styles, leaving cross-speaker robustness largely unexplored.

By isolating character subsets, we provide a controlled analysis of delivery *intent* and *style*, demonstrating that expressive prosody and gestural cues lead to significantly higher multimodal recognition accuracy. This focus on personality-aware evaluation introduces a new dimension to multimodal sarcasm understanding.

2.4 Positioning of This Work

Prior research has progressively expanded from text-only sarcasm modeling to large multimodal architectures emphasizing incongruity learning and adaptive fusion. However, these systems often trade interpretability for complexity. Our framework contributes a complementary perspective: a lightweight model that uses fully-frozen pretrained embeddings to systematically disentangle modality contributions and evaluate how delivery style influences multimodal detectability. By combining reproducibility with efficiency, we offer a transparent baseline for future multimodal sarcasm research.

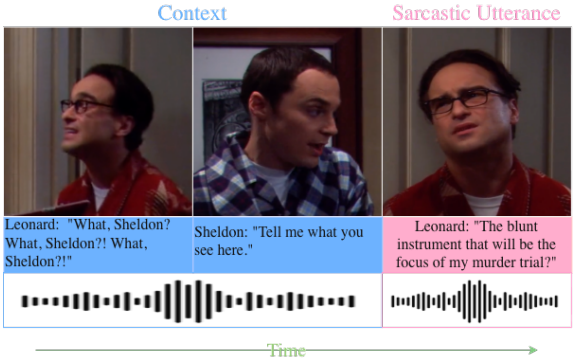

Figure 1: Example from the MUStARD++ Balanced dataset. Each instance consists of a dialogue context and a target utterance aligned across text, audio, and video modalities.

3 Dataset

3.1 MUStARD++ Balanced Overview

All experiments in this work are conducted on the MUStARD++ Balanced dataset (Bhosale et al., 2023), a curated and class-balanced variant of the MUStARD++ corpus. The MUStARD dataset (Castro et al., 2019) originally introduced multimodal sarcasm detection using aligned text, audio, and video segments from television sitcoms. MUStARD++ (Ray et al., 2022) extended this resource with additional clips, improved alignment, richer annotations, and emotion labels, enabling more detailed analysis of sarcasm expression. MUStARD++ Balanced further refines the corpus by addressing label imbalance and removing samples with unreliable visual signals, resulting in a more stable benchmark for multimodal learning. This version is now commonly used in recent work on multimodal sarcasm detection and allows for controlled comparison across modalities without confounding effects from skewed class distributions.

3.2 Data Composition

MUStARD++ Balanced includes thousands of annotated utterances, with each entry aligned across text, audio, and visual modalities. For this work,

we focus on the utterance-level subset containing clips with full multimodal alignment, i.e., where both speech and facial frames are synchronized with the transcribed text. This subset ensures consistent modality availability for all samples, without relying on missing-modality imputation or augmentation. Metadata from MUStARD++ Balanced's extended annotation file enables grouping by character, allowing the analyses in Sec. 5.5.

4 Methodology

This work aims to evaluate how far multimodal sarcasm understanding can be achieved using lightweight architectures built entirely from frozen pretrained embeddings. Our framework combines textual, acoustic, and visual cues extracted from the MUStARD++ Balanced dataset, emphasizing efficiency, interpretability, and robustness over heavy fine-tuning. Fig. 2 provides a high-level overview.

4.1 Preprocessing and Feature Extraction

Raw videos were separated into two groups: *utterance_videos* and *utterance_additions*, each containing short contextual segments. Speech was extracted and resampled to 16 kHz, producing the `audio_wav16k` directory used for acoustic embedding generation via wav2vec 2.0. Visual features were derived from frame-level OpenFace outputs, including facial Action Units (AUs), head pose, and gaze direction. All features were merged into a unified "wide" representation, consisting of 3,190 multimodal samples.

Text Each utterance and its immediate context are tokenized and encoded using the pretrained `RoBERTa-base` (Liu et al., 2020) model from Hugging Face Transformers. We use the pooled `[CLS]` representation (768 dimensions) as a fixed textual embedding for each utterance.

Audio All speech clips are resampled to 16 kHz and processed with Baevski et al. (2020)'s pretrained wav2vec 2.0 encoder. Frame-level outputs are mean-pooled to form a 768-dimensional vector capturing prosodic patterns such as pitch, energy, and rhythm—key indicators of sarcastic tone.

Visual Each video segment is analyzed using OpenFace 2.0 (Baltrusaitis et al., 2018) to extract facial AUs, head pose, and gaze direction. For each AU and geometric feature, we compute statistical descriptors (mean, standard deviation, range, and slope) over time, resulting in a 1,800-dimensional visual feature vector per utterance.

Character	Utterances	Non-Sarcastic	Sarcastic
Chandler	156	38	118
Sheldon	126	65	61
House	137	62	75
Howard	136	68	68
Penny	125	54	71
Leonard	110	52	58

Table 2: Speaker-wise sarcasm distribution. Only characters with at least 100 utterances are included to ensure stable evaluation.

4.2 Fusion Strategies

To examine the interaction between modalities, three fusion strategies are explored: 1) **Early Fusion** concatenates embeddings from all active modalities and passes them through a shallow feedforward layer or logistic regression classifier; 2) **Late Fusion** trains separate unimodal classifiers and combines their prediction probabilities through weighted averaging or meta-classification; 3) **Stacking Fusion** uses intermediate unimodal representations as inputs to a secondary classifier, allowing limited cross-modal interaction while retaining modality specialization.

All encoders remain frozen during training, ensuring that observed differences stem from differences in fusion rather than in representations.

Fusion Heads For each fusion strategy, we evaluated two compact classifiers: (1) a *logistic regression* (LR) head, which performs linear combination of modality embeddings, and (2) a shallow *multilayer perceptron* (MLP) head consisting of a Dense–ReLU–Dropout–Linear stack (approximately 0.5M parameters). Both heads operate on frozen pretrained embeddings without end-to-end fine-tuning. Unless otherwise specified, all subsequent analyses and ablations use the LR variant, which consistently provided higher and more stable performance across folds.

4.3 Speaker Filtering and Subsets

To explore how personality and speaking style influence sarcasm expression, we construct speaker-specific subsets. Speaker identity strongly conditions the style of sarcasm expression. For instance, Sheldon Cooper's (*TBBT*) sarcastic utterances primarily exhibit *unintentional sarcasm*, characterized by literal tone and minimal prosodic variation, whereas Chandler Bing's (*Friends*) sarcasm is overt and expressive. After mapping video identifiers to annotation keys, we selected speakers who have more than 100 utterances in the dataset, forming the basis for our character-wise experiments. Utter-

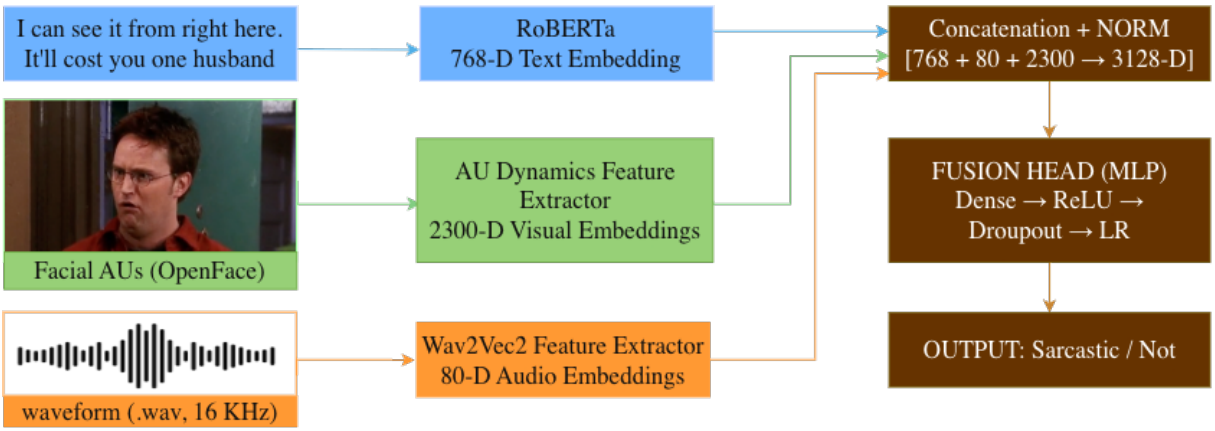

Figure 2: Lightweight multimodal sarcasm detection pipeline. RoBERTa, Wav2Vec2, and OpenFace extract frozen text, audio, and visual embeddings, which are concatenated into a unified representation and passed through a shallow MLP fusion head to predict sarcasm labels.

ance counts are given in Table 2.

For character-wise experiments, each speaker subset is evaluated independently using the same fusion configurations as the full-dataset experiments. This setup isolates how sarcasm delivery style affects multimodal detectability without introducing cross-speaker confounds.

4.4 Training Protocol

Experiments are conducted using five-fold grouped cross-validation, where utterances from the same dialogue segment never co-occur in both training and testing sets. This strategy prevents context leakage and mirrors natural discourse boundaries. Models are optimized using the Adam optimizer with a learning rate of 1×10^{-4}, batch size of 16, and early stopping based on validation F1-score. Macro-F1 is used as the primary evaluation metric to account for class imbalance.

Additionally, we conduct *leave-one-speaker-out* (LOSO) evaluations to assess cross-speaker generalization. Here, all utterances from a held-out speaker are used exclusively for testing, while models are trained on the remaining speakers. LOSO experiments are performed for *Sheldon Cooper*, *Chandler Bing*, and *Dr. Gregory House*—whose sarcasm style is dry and deadpan, similar to Sheldon's—enabling direct comparison between in-domain and out-of-speaker performance.

4.5 Lightweight Implementation

The entire pipeline is designed for computational efficiency. Across all configurations, fewer than 1M trainable parameters are used, well under 0.1% of typical fine-tuned transformer models (see Table 1). Training followed a five-fold **stratified group cross-validation** protocol, ensuring that utterances from the same dialogue segment never appeared in both train and test splits. See Appendix A for further experimental details.

5 Results and Analysis

This section presents both the quantitative and qualitative evaluation of the unimodal and multimodal sarcasm detection models on the MUStARD++ Balanced dataset. All experiments were conducted using five-fold stratified group cross-validation, ensuring that utterances from the same dialogue segment never appear in both training and test splits. Each evaluation was run 3 times with different random seeds, for a total of 15 runs. Performance is reported in terms of macro-F1 score.

5.1 Overview

Our primary objective is not to achieve state-of-the-art performance, but to demonstrate that lightweight architectures can make effective use of pretrained embeddings to achieve meaningful multimodal sarcasm understanding without large-scale fine-tuning. All models in this study rely on frozen, pretrained embeddings for text, audio, and visual modalities, combined through compact fusion heads such as logistic regression or shallow MLPs. This setup isolates the contribution of each modality and fusion strategy, removing the confounding influence of model size or fine-tuning.

Although absolute F1 scores may match but rarely exceed those of fine-tuned Transformers, they reveal consistent and interpretable patterns. Text–audio combinations consistently outperform unimodal variants, and early or stacking fusion yields stronger generalization than late fusion.

Modality	Mean ± SD (F1)
Text (RoBERTa)	0.64 ± 0.03
Audio (Wav2Vec2.0)	0.59 ± 0.04
Visual Dynamics (OpenFace)	0.54 ± 0.02

Table 3: Unimodal macro-F1 results on MUStARD++ Balanced using five-fold cross-validation.

Moreover, speaker-specific analyses indicate that the presence or absence of expressive multimodal cues substantially affects detection accuracy. These findings reinforce our central claim: that sarcasm recognition depends more on *how* modalities interact—and on *how* sarcasm is expressed—than on the scale of the underlying models.

5.2 Unimodal Performance

Table 3 summarizes the unimodal baselines. All models rely on frozen, pretrained embeddings for each modality, preserving the lightweight setup described earlier. Among single modalities, the acoustic model performs better than the visual model, confirming that intonation and prosody encode sarcasm more reliably than facial dynamics in this dataset. This aligns with prior observations that sarcastic tone often carries stronger discriminative cues than subtle or inconsistent facial expressions, particularly across television dialogue. The textual model remains the strongest unimodal baseline, capturing the linguistic contrasts and contextual markers that often signal sarcastic intent. These unimodal results establish a foundation for analyzing how cross-modal fusion amplifies or, in some cases, fails to amplify—sarcasm-specific signals.

5.3 Fusion Strategies

We evaluate multimodal sarcasm detection using three fusion strategies: early fusion, late fusion, and stacking—across two classifiers: Logistic Regression (LR) and a multilayer perceptron (MLP). All models use frozen pretrained embeddings, and performance is reported as mean Test F1 with standard deviation across folds, as shown in Table 4.

Early fusion with LR achieves the strongest overall performance, obtaining an F1 score of 0.68 ± 0.07 when combining text and audio features. This configuration also exhibits the highest variance, indicating sensitivity to data splits despite strong average performance. Adding visual features in the early fusion setting does not improve results for LR and instead reduces performance to 0.60 ± 0.01.

LR with late and stacking fusion yield lower but more stable performance compared to early fusion. Late fusion achieves 0.63 ± 0.05 with text and audio and drops to 0.59 ± 0.02 with visual features.

Classifier	Fusion Strategy	T + A	T + A + V
LR	Early Fusion	**0.68 ± 0.07**	0.60 ± 0.01
LR	Late Fusion	0.63 ± 0.05	0.59 ± 0.02
LR	Stacking	0.61 ± 0.01	0.59 ± 0.02
MLP	Early Fusion	0.59 ± 0.03	0.57 ± 0.03
MLP	Late Fusion	0.61 ± 0.02	0.59 ± 0.04
MLP	Stacking	0.63 ± 0.01	0.63 ± 0.01

Table 4: Comparison of fusion strategies on MUStARD++ Balanced using frozen pretrained embeddings. Early fusion with LR yields the best and most stable performance across folds.

A similar pattern is observed for stacking, where performance remains comparable across feature combinations but does not surpass early fusion.

MLP-based models demonstrate more consistent behavior across fusion strategies. The early fusion MLP reaches 0.59 ± 0.03 for text and audio, with a slight decrease when visual features are added. Late fusion improves modestly over early fusion for text and audio (0.61 ± 0.02), but again declines with the inclusion of visual features.

Stacking with MLP provides the most balanced performance, at 0.63 ± 0.01 for both text+audio and text+audio+visual. While this approach does not outperform LR early fusion with text+audio, it offers improved stability and robustness.

Overall, these results indicate that while multimodal fusion improves over unimodal baselines, the inclusion of visual features does not consistently yield gains and, in several cases, slightly degrades performance. Text and audio remain the dominant contributors to sarcasm detection performance in this setting.

5.4 Comparison with Prior Work

Table 1 showed the performance of recent multimodal sarcasm detection systems evaluated on MUStARD++ or closely related datasets. While large transformer-based models (e.g., MSAF-SDM) achieve higher absolute F1 scores, they involve tens to hundreds of millions of parameters and require fine-tuning. With **fewer than 1 million trainable parameters**, our frozen-feature fusion approach attains performance that either exceeds or reaches to within 0.05 F1 of prior approaches (e.g., Jang and Frassinelli (2024), Phukan et al. (2024), or Bhosale et al. (2023)[1]), illustrating that interpretability and efficiency need not come at the expense of multimodal understanding.

[1]We note that Jang and Frassinelli (2024) and Phukan et al. (2024) evaluate on MUStARD++, not MUStARD++ Balanced.

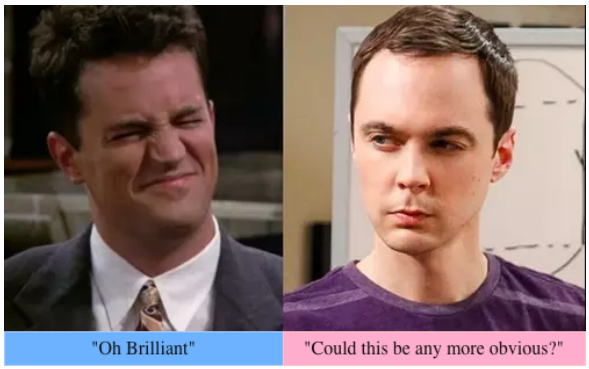

Figure 3: Illustrative examples of sarcasm styles in MUStARD++ Balanced: Sheldon (right) shows subtle, context-dependent irony, while Chandler (left) exhibits overt, deliberate sarcasm.

Character	Mean± SD (F1)	Text (%)	Audio (%)	Visual (%)
Chandler	**0.68 ± 0.07**	29.52	24.13	**46.34**
Sheldon	0.44 ± 0.08	35.70	16.13	**48.17**
House	0.44 ± 0.14	29.42	25.95	**44.63**
Howard	0.58 ± 0.10	35.73	19.23	**45.04**
Penny	0.61 ± 0.07	25.93	19.54	**54.52**
Leonard	0.47 ± 0.03	27.80	22.77	**49.43**

Table 5: Speaker-wise sarcasm detection performance and relative modality contribution (%) based on normalized absolute weights from the early-fusion LR model.

5.5 Speaker-wise Analysis: Intentional vs. Unintentional Sarcasm

To better understand the role of multimodal expressive cues in sarcasm perception (e.g., Fig. 3), we conducted a speaker-specific analysis across characters, using the splits described in Table 2, with identical features and training configurations.

As shown in Table 5, characters with overt and expressive delivery styles, such as Chandler and Penny, achieve markedly higher macro-F1 scores than speakers whose sarcasm is more subtle or context-dependent. In contrast, Sheldon and Dr. House exhibit significantly lower performance. These speakers frequently deliver sarcastic remarks with restrained prosodic variation or facial expression, making it harder to distinguish sarcasm from literal speech. Howard Wolowitz and Leonard Hofstadter fall between these extremes, with moderate detectability consistent with their mixed expressive styles, including both overt and deadpan delivery.

Importantly, these differences emerge under a controlled experimental setup, where each speaker is evaluated independently using the same multimodal features and classifier architecture. This suggests that variability in sarcasm recognition is driven primarily by speaker delivery characteristics rather than differences in data volume or class balance. Together, these results highlight

Speaker	LR		MLP	
	T + A	T + A + V	T + A	T + A + V
Sheldon	0.52	0.51	0.63	0.59
House	0.56	0.55	0.50	0.58
Chandler	0.60	0.63	0.62	0.62

Table 6: Leave-one-speaker-out (LOSO) results comparing Logistic Regression and MLP classifiers.

that sarcasm detectability in multimodal systems is strongly speaker-dependent, motivating explicit consideration of delivery style in model evaluation. *This analysis constitutes a core contribution of this paper*: no previous work, including those that developed large SOTA transformer approaches, have investigated robustness and the relative contribution of different modalities to different speakers and delivery styles.

To further interpret character-level differences, we analyzed the weights of the best-performing early-fusion LR model to estimate the relative contribution of each modality for each speaker. Across all characters, visual features account for the largest share of model weight, followed by textual features, with audio contributing a smaller portion (Table 5). Characters with more expressive delivery styles (Chandler, Penny), show a stronger reliance on visual cues, while characters with flatter or more monotone delivery (Sheldon, House), exhibit relatively higher dependence on textual information.

These results indicate that modality contributions are speaker-dependent: while visual features consistently influence the model's decisions, their effectiveness varies by character, reinforcing the need for speaker-aware analysis when interpreting multimodal sarcasm detection models.

Leave-One-Speaker-Out (LOSO) Evaluation To evaluate cross-speaker generalization, we conduct leave-one-speaker-out (LOSO) experiments for three high-frequency speakers: *Sheldon Cooper*, *Dr. Gregory House*, and *Chandler Bing*. In each setting, all utterances from the target speaker are excluded from training and used exclusively for testing, while models are trained on all remaining speakers. All experiments use the same frozen features and fusion configurations as earlier sections.

Across all three speakers, LOSO performance is lower than the corresponding in-domain evaluations, indicating that exposure to samples from specific speakers contributes to multimodal sarcasm detection performance (Table 6) Results vary by speaker and model configuration, with no single best-performing classifier-modality combination.

Speaker	LR		MLP	
	T + A	T + A + V	T + A	T + A + V
Sheldon	0.58	0.54	0.64	0.67
House	0.61	0.61	0.65	0.62

Table 7: Ablation test results on Sheldon/House samples when excluding both Sheldon and House from training.

Cross-Speaker Ablation To further isolate cross-speaker effects, we perform ablation-style transfer experiments by jointly removing two speakers with qualitatively similar deadpan delivery styles: *Sheldon Cooper* and *Dr. Gregory House*, from training and evaluating on each speaker independently. This setting tests whether models trained without exposure to either speaker can generalize to their sarcasm styles when both are excluded from the training distribution. We report both the standard LOSO results and the cross-speaker ablation results using identical model configurations.

For both speakers, performance under the cross-speaker ablation setting differs from standard LOSO evaluation (Table 7), demonstrating that model behavior depends not only on whether a speaker is held out, but also on which other speakers are present during training.

5.6 Discussion and Error Analysis

The results demonstrate that prosodic information complements textual context more effectively than visual features, which often introduce noise or inconsistency across speakers. In our experiments, adding visual features rarely improved performance and occasionally degraded it, particularly under MLP and late fusion configurations. This is despite visual features being allocated a high weight by a logistic regressor, and aligns with unimodal and fusion results where visual consistently ranked lowest in standalone performance—meaning that when they are included, visual features are weighted heavily but contain inconsistent information.

The LOSO experiments further highlight the role of speaker identity: all three held-out speakers exhibited reduced performance relative to their in-domain evaluations, even under identical feature and classifier setups. This suggests that sarcasm detection is not merely a function of delivery modality but is sensitive to speaker-specific patterns. For example, *Chandler* retained high performance under LOSO (up to 0.63 F1 with LR + TAV), while *Sheldon* dropped substantially (as low as 0.51). This again reflects that overt sarcasm transfers more robustly than subtle or monotone delivery styles.

Ablation tests reinforced this pattern. When both *House* and *Sheldon* were removed from training and tested individually, performance on each actually improved over standard LOSO in several configurations, particularly for MLP + TAV, where *Sheldon* reached 0.67 and *House* reached 0.62. This suggests that models may overfit to misleading speaker-specific cues when a speaker is present during training, or that certain speaker combinations interfere with generalization.

These results underscore that model robustness in sarcasm detection depends heavily on speaker composition, not just modality alignment or classifier complexity. The findings argue for evaluating models in speaker-exclusion settings when claiming generalization, and for future work to explore speaker-invariant representations.

6 Conclusion

This paper investigated multimodal sarcasm detection using frozen pretrained embeddings across text, audio, and visual modalities, evaluated on the MUStARD++ Balanced dataset. Our baseline experiments confirmed that textual and prosodic features outperform visual features in both unimodal and fusion settings. Visual cues, as represented in this dataset, contributed little to overall performance and, in some cases, degraded results, particularly under MLP models and late fusion.

Through extended speaker-wise analysis, we evaluated performance across six high-data characters. Results revealed substantial variation: some speakers (e.g., *Chandler*) achieved high scores across all modalities, while others (e.g., *Sheldon*, *House*) performed poorly, even with all three modalities combined. These trends were confirmed through modality contribution analysis using LR coefficient weights.

To test model generalizability, we conducted leave-one-speaker-out (LOSO) evaluations for three characters and cross-speaker ablation tests where two speakers were excluded from training entirely. LOSO consistently yielded lower performance compared to in-domain results, confirming a degree of speaker overfitting. Interestingly, ablation experiments showed that excluding certain speakers during training sometimes improved performance on them, indicating interference effects or misleading speaker-specific patterns.

In sum, our results suggest that speaker identity remains a major challenge for robust multimodal sarcasm detection. Future work should emphasize

speaker-invariant modeling, dynamic fusion strategies, and better exploitation of visual cues, especially where new or unseen speakers are common.

Limitations

Although this study provides a reproducible and efficient baseline for multimodal sarcasm detection, several limitations remain. First, MUStARD++ Balanced, while consisting of a diverse sample of TV shows, is limited to scripted television dialogue, which may not generalize to spontaneous or cross-cultural sarcasm. Second, our analysis relies on frozen pretrained encoders, which constrain modality adaptation and may underrepresent subtle expressive nuances. Third, visual data quality varies substantially across clips, and missing or low-resolution facial cues can weaken multimodal consistency. Finally, we focus on two speakers for controlled analysis; extending this approach to broader conversational or multilingual settings would strengthen ecological validity and generalizability.

Ethical and Reproducibility Notes

Our experiments are conducted over publicly-available data from television shows, and so as also mentioned in Limitations, methods for this domain may not generalize to spontaneous conversation or settings with different cultural norms, and automatic classification of phenomena such as sarcasm should be treated cautiously in real-life situations where it may be misinterpreted or lead to misunderstanding.

All media in MUStARD++ Balanced are publicly available under fair-use research provisions. To ensure reproducibility, we use only official annotations and extracted features without altering dialogue content. A link to our code is provided in Sec. 1. By relying on frozen pretrained encoders and lightweight fusion heads, the full experimental pipeline can be reproduced without access to specialized hardware or large-scale distributed training resources.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.

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. *arXiv preprint arXiv:2310.01430*.

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.

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.

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

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! 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 4238–4249.

Jiby Jose. 2025. An efficient sarcasm detection in audio using parameter-reduced depthwise cnn. *International Journal of Innovative Research in Advanced Engineering*, 12.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Arpan Phukan, Santanu Pal, and Asif Ekbal. 2024. Hybrid quantum-classical neural network for multimodal multitask sarcasm, emotion, and sentiment analysis. *IEEE Transactions on Computational Social Systems*, 11(5):5740–5750.

Devraj Raghuvanshi, Xiyuan Gao, Zhu Li, Shubhi Bansal, Matt Coler, Nagendra Kumar, and Shekhar Nayak. 2025. Intra-modal relation and emotional incongruity learning using graph attention networks

for multimodal sarcasm detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. 2022. A multimodal corpus for emotion recognition in sarcasm. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 3486–3493, Marseille, France.

Zihang Wu and Jiali Zang. 2024. Multi-scale adaptive fusion with shared discrepancy minimization for multimodal sarcasm detection. *Knowledge-Based Systems*, 293:111715.

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2024. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *arXiv preprint arXiv:2410.18882*.

Bingzhe Zhou, Hannan Wang, Yuan Yao, Taolue Chen, Feng Xu, and Xiaoxing Ma. 2024. Simulate, refine and integrate: Strategy synthesis for efficient smt solving. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page 7976–7984. International Joint Conferences on Artificial Intelligence Organization.

A Additional Experimental Details

All experiments were conducted using Python 3.10 and PyTorch 2.1 within a dedicated environment. Experiments were run on standard research GPUs without requiring large-scale distributed training. The frozen-encoder design significantly reduces both memory usage and training time relative to fully fine-tuned multimodal architectures. Frozen pretrained encoders were used for each modality: RoBERTa-base for text, wav2vec 2.0 for audio, and OpenFace-derived Action Unit dynamics for visual features. Embeddings from each modality were standardized and fused through lightweight classifiers (logistic regression or shallow MLPs) implemented in scikit-learn.

Training followed a five-fold **stratified group cross-validation** protocol, ensuring that utterances from the same dialogue segment never appeared in both train and test splits. Each fold used a batch size of 16, the Adam optimizer with a learning rate of 1e–4, and early stopping based on validation F1. All cross-validation results report the mean and standard deviation of the macro-F1 across the five folds, whereas the leave-one-speaker-out (LOSO) and cross-speaker ablation experiments report results from single runs.

B Reproducibility and Consistency Check

To validate robustness, we repeated the early-fusion experiments described in Sec. 5.3 across multiple random seeds using the same grouped cross-validation protocol. While absolute F1 values varied slightly (typically within ± 0.05), the relative performance trends remained consistent: *(i)* text–audio consistently outperformed single-modality models, and *(ii)* adding visual dynamics did not yield further gains. These observations reinforce the stability of our lightweight fusion architecture across runs.