Knowledge-Based Multimodal and Context Incongrity Modeling For Sarcasm Detection

Wenjie Xu, Zhong Qian*, Peifeng Li, Qiaoming Zhu
School of Computer Science and Technology, Soochow University, Suzhou, China
20245227044@stu.suda.edu.cn, {qianzhong, pfli, qmzhu}@suda.edu.cn

***Abstract*—Multimodal sarcasm detection is a challenging problem in sentiment analysis, which requires recognizing mismatches between literal content and intended meaning. While recent methods have achieved promising results, they often emphasize inconsistencies associated with strongly affective words, potentially neglecting informative signals conveyed by other tokens. Moreover, without adequate commonsense knowledge, existing approaches may not fully utilize dialogue context, resulting in degraded performance on cases that require deeper semantic inference. To address these limitations, we propose a framework that jointly models fine-grained cross-modal conflicts and knowledge-enhanced contextual semantics. Specifically, we develop a Token-level Multimodal Inconsistency Modeling module to capture token-wise contradictions between textual and non-verbal cues. In addition, we introduce a Knowledge-based Contextual Semantic Inconsistency Modeling module that leverages large language models (LLMs) to incorporate external commonsense, facilitating the identification of implicit logical inconsistencies within the dialogue context. Experiments on the MUStARD benchmark show that our approach yields consistent improvements over prior work. Our model achieves an F1 score of 77.9 in the speaker-dependent setting and maintains competitive performance in the speaker-independent setting.**

***Index Terms*—multimodal, sarcasm detection**

I. Introduction

Sarcasm is a sophisticated linguistic phenomenon where the intended meaning is opposite to the literal expression. With the rise of social media, sarcasm has become ubiquitous in multimodal data, posing significant challenges to sentiment analysis systems. Multimodal Sarcasm Detection, which leverages textual, visual, and acoustic cues, has emerged as a critical research field.

According to [1], sarcasm can be categorized into distinct types. For instance, Illocutionary Sarcasm relies heavily on non-verbal cues, whereas Propositional Sarcasm demands context information for precise detection. This suggests that an effective sarcasm detection model must simultaneously handling inconsistencies in non-verbal behaviors and logical inconsistencies within the context.

Although recent advancements have transitioned from text-only analysis to multimodal fusion, existing methods face limitations in addressing the aforementioned complexities. On one hand, regarding Illocutionary Sarcasm, while previous studies [2], [3] attempt to capture inconsistencies, they predominantly focus on words with positive sentiment. This selective attention overlooks the fact that neutral or functional words can also trigger sarcasm when paired with specific tones or micro-expressions. On the other hand, for Propositional Sarcasm,

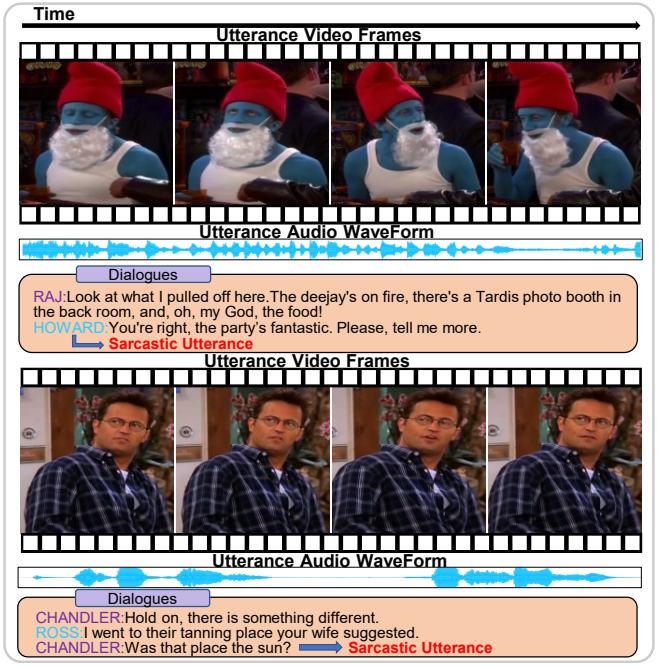

Fig. 1. Cross-modal Incongruity Instance(top), contextual Incongruity Instance(bottom).

existing multimodal models perform poorly. Although some studies [2], [4] attempt to incorporate contextual history, the results show performance degradation or marginal gains. We argue that this is primarily due to a deficiency in commonsense background knowledge.

We illustrate our motivation using Fig 1. In Fig 1 (top), the current utterance aligns with the overall tone of the preceding context. However, a conflict arises within the multimodal channels. While the literal text conveys strong interest and agreement, the non-verbal cues convey the opposite message. Specifically, 'fantastic' is paired with an eye-roll, and 'tell me more' is accompanied by a flat tone. Creating inconsistency between the enthusiastic literal meaning and the indifferent non-verbal signals. Conversely, in Fig 1 (bottom), the speaker maintains a serious demeanor, exhibiting no intra-modal inconsistency. However, by connecting the previously mentioned 'tanning place' with 'the sun' in the current utterance, and leveraging a LLMs to supplement related concepts (e.g., skin care effects), one can identify that the speaker employs extreme hyperbole to mock the other person's tanning results. This demonstrates that detecting such sarcasm requires external knowledge supplementation and logical reasoning re-

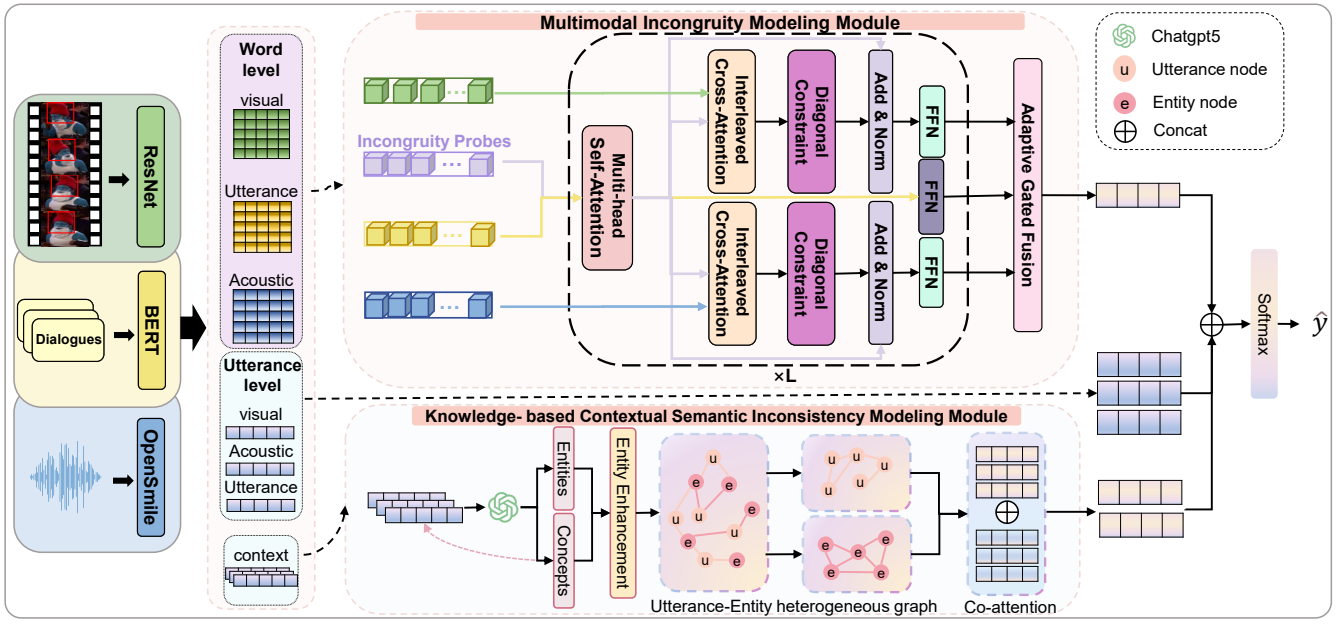

Fig. 2. The overall structure of our model.

garding contextual semantics, rather than superficial behavioral cues.

To address these limitations, we introduce a Word-level Multimodal Inconsistency Modeling Module, utilizing learnable probes to capture cross-modal inconsistencies. Furthermore, we propose a Knowledge-based Contextual Semantic Inconsistency Modeling Module, which leverages LLMs to introduce additional knowledge and utilizes heterogeneous graphs to model inconsistencies within the context. Experiments on the MUStARD dataset demonstrate the effectiveness of our method, achieving an F1-score of 77.9 with a 1.8% improvement. Our main contributions are as follows:

- We propose a Multimodal Inconsistency Modeling Module that captures contradictions across all words, transcending the limitation of focusing on sentiment-bearing words.
- We design a Knowledge-based Contextual Semantic Inconsistency Mining Module. We utilize LLMs to supplement commonsense and construct heterogeneous graphs, enabling the model to reveal semantic logical disconnects between the current utterance and dialogue context.
- Extensive experiments conducted on the MUStARD dataset show that our framework outperforms state-of-the-art methods in both speaker-dependent and speaker-independent settings, verifying the effectiveness and robustness of our approach.

II. Related Work

A. Text-based Sarcasm Detection

Early research on sarcasm detection primarily relied on rule-based techniques and manual feature engineering. These methods often viewed sarcasm as a linguistic incongruity between positive sentiment and negative situations [5]–[7]. With the advent of deep learning, researchers adopted neural architectures, such as CNNs, RNNs, and Graph Neural Networks (GNNs), to automatically learn features and capture intra-textual contradictions [8], [9]. Furthermore, some studies recognized the importance of context, incorporating speaker history and personality traits to aid judgment [10]. However, linguistic studies suggest that sarcasm relies heavily on paralinguistic cues (e.g., intonation, facial expressions), implying that text-only approaches possess inherent limitations [11].

B. Multimodal Sarcasm Detection

To address these limitations, Multimodal Sarcasm Detection integrates textual, visual, and acoustic cues. The release of the MUStARD dataset [4] established a standard benchmark for trimodal analysis. While early approaches utilized straightforward feature fusion strategies [4], subsequent research focused on designing complex attention mechanisms to model interactive relationships between modalities [12].

Recent methods prioritize capturing cross-modal *inconsistency*. For instance, [2] proposed an Inconsistency-Aware Attention Network (IWAN) to score word-level contradictions between text and non-verbal signals. Building on this, [3] introduced a context-aware self-attention framework that integrates instantaneous multimodal information. They further employed a word weighting mechanism to specifically highlight discrepancies between the literal meaning of positive sentiment words and the connotation derived after multimodal fusion. Reference [13] leveraged GATs for intra-modal relations and contrastive attention to capture cross-modal emotional incongruities. While effective, these methods largely focus on explicit emotional vocabulary, often overlooking subtle cues present in other tokens.

III. METHOD

We present a framework designed to capture both fine-grained cross-modal conflicts and deep contextual incongruities, as illustrated in Fig 2. We first extract multimodal

Prompt:
You are building a knowledge graph. Your task is to link and conceptualize entities from the given text. Please extract it step by step:

(1) Extract the entities in the given text. Entities can be nouns or noun phrases that refer to real-world objects or abstract concepts. You can use named entity recognition (NER) tools or part-ofspeech (POS) taggers to identify entities.

(2) Link the extracted entities to the existing knowledge base or concept system. By comparing and matching the entity's context, entity attributes, etc., it is associated with the corresponding entity in the knowledge base, and the link between the entity and the knowledge base is established.

(3) For each identified entity, we obtain its concept information from the existing knowledge graph through conceptualization, adopting the "isA" relationship to obtain the concept. Conceptualization: Map the extracted entities to related concepts or topics. This step associates entities with a set of concepts or topics in order to better understand the semantic meaning and domain of entities.

(4) Most importantly, if you can't find any knowledge, output: "None".

You need to generate the following result:

Entity Set T= {'Entity 1', 'Entity 2', 'Entity 3', ..., 'Entity N'}
The set of concepts corresponding to each entity:
C'Entity 1' = {'Concept 1', 'Concept 2', 'Concept 3', ..., 'Concept N'}
C'Entity 2' = {'Concept 1', 'Concept 2', 'Concept 3', ..., 'Concept N'}
C'Entity N' = {'Concept 1', 'Concept 2', 'Concept 3', ..., 'Concept N'}

Fig. 3. Prompt used in the ChatGPT-based Knowledge Enhancement module

features and then feed them into our model, which consists of two key components: (1) Multimodal Inconsistency Modeling, which utilizes a learnable probe matrix inspired by Q-former [14] to scrutinize discrepancies between textual and non-verbal modalities; (2) Knowledge-based Contextual Inconsistency Modeling, which leverages external knowledge and heterogeneous graphs to uncover logical breaks between the dialogue context and the current utterance. Finally, features from these modules are adaptively fused for sarcasm prediction.

A. Multimodal Feature Extraction

Utterance-level Features. Following previous work [4], we extract global textual (t_u), visual (v_u), and acoustic (a_u) features for the entire utterance using pre-trained BERT, ResNet50, and OpenSmile, respectively.

Word-level Features. Since sarcasm often relies on token-wise cues easily overlooked by global features, we extract fine-grained word-level representations. First, we employ the GENTLE forced aligner to synchronize the audio and text, obtaining precise timestamps for each word sequence $\{w_1, w_2, ..., w_n\}$. For visual modality, we focus on the speaker's facial expressions to reduce background noise. We detect faces via MTCNN [15] and perform active speaker identification using Inception ResnetV1 [16]. Visual features for each word, denoted as v_{wi}, are then extracted via ResNet50 [17] by averaging frame-level features within the corresponding timestamps. For textual modality, we utilize BERT [18] to obtain word-level embeddings t_{wi} from the last hidden layer. For acoustic modality, we use OpenSmile to extract Low-Level Descriptors such as MFCCs and pitch for each aligned segment, averaging them to form acoustic features a_{wi}.

B. Multimodal Inconsistency Modeling

Sarcasm often manifests as conflicts between literal text and non-verbal cues. To capture these fine-grained contradictions, we propose a Word-level Multimodal Inconsistency Modeling module. We introduce learnable probes aligned with text tokens to 'scrutinize' visual and acoustic modalities.Through a diagonal attention constraint, the model focuses exclusively on cross-modal discrepancies at the same token position.

Given aligned features H_t, H_v, H_a, we first project them into a shared space via linear layers to obtain $X_t, X_v, X_a \in \mathbb{R}^{n\times d}$,where d is the number of hidden units. Next, we introduce a learnable probe matrix $P \in \mathbb{R}^{n\times d}$ corresponding one-to-one with the token positions, initialized using pre-trained weights from BERT [18]. The probes are updated layer-by-layer through L stacked modules. At the ℓ-th layer, we first feed P and X_t into a parameter-shared multi-head self-attention(MSA) module, enabling each probe to absorb global textual semantics while maintaining positional correspondence. This yields text-aware representations that serve as semantic anchors for cross-modal inconsistency probing:

$$[\bar{P}^{\ell}; X_t^{\ell}] = MSA([\bar{P}^{(\ell-1)}; X_t^{(\ell-1)}]) \tag{1}$$

To efficiently mine inconsistencies, we employ an Interleaved Cross-Attention strategy, activating cross-modal probing every k layers. When active, the probes $\bar{P}^{\ell}$ interact with non-verbal features $X_m \in \mathbb{R}^{n\times d}$,$m \in \{v, a\}$. We impose a Diagonal Constraint to restrict attention to the same token position. For the i-th head, the attention score S_i^m is computed as the diagonal of the standard attention matrix and weights the value vector element-wise :

$$S_i^m = \text{diag}\left(\frac{(\bar{P}^{(\ell)}W_q^i)(X_m W_{k,m}^i)^T}{\sqrt{d_k}}\right) \tag{2}$$

$$Head_i^m = \text{Softmax}(S_i^m) \odot (X_m W_{v,m}^i) \tag{3}$$

where S_i denotes the cross-modal inconsistency for each token in the i-th subspace, $W_q^i, W_{k,m}^i, W_{v,m}^i$ are learnable parameters, $d_k = d/I$, and I is the number of attention heads.

The outputs of all heads are concatenated and fused via a linear layer W_o, followed by a residual connection to obtain the output of that layer:

$$\Delta Q_m^{(\ell)} = Concat(Head_1^m, Head_2^m, ..., Head_i^m)W_o \tag{4}$$

$$Q_m^{(\ell)} = \bar{Q}^{(\ell)} + \Delta Q_m^{(\ell)} \tag{5}$$

where $W_o \in \mathbb{R}^{d\times d}$. When cross-modal probing is not activated at layer ℓ, we simply set $Q_m^{(\ell)} = \bar{Q}^{(\ell)}$.

Finally, the inconsistency feature $Q_m^{(\ell)}$ and the refined textual feature X_t^{ℓ} are fed into their respective feed-forward networks before entering the next layer. After L layers of interaction, we obtain the visual inconsistency feature $Q_v^{(L)}$, the acoustic inconsistency feature $Q_a^{(L)}$, and the interacted textual feature X_t^L. To integrate information from different modalities, we propose an adaptive gating fusion mechanism to generate the final multimodal inconsistency feature H_m:

$$g_{tv} = \sigma\left([X_t^L \oplus Q_v^{(L)}]W_{tv} + b_{tv}\right) \tag{6}$$

$$g_{ta} = \sigma\left([X_t^L \oplus Q_a^{(L)}]W_{ta} + b_{ta}\right) \tag{7}$$

$$H_{tv} = X_t^L + g_{tv} \odot Q_v^{(L)} \tag{8}$$

$$H_{ta} = X_t^L + g_{ta} \odot Q_a^{(L)} \tag{9}$$

$$H_{tva} = X_t^L + g_{tv} \odot Q_v^{(L)} + g_{ta} \odot Q_a^{(L)} \tag{10}$$

where $\sigma(\cdot)$ is the Sigmoid function, $b_{tv}, b_{ta} \in \mathbb{R}^{n\times 1}$, and W_{tv}, W_{ta} are learnable parameters. Finally, we aggregate information from different fusion paths and perform summation

TABLE I
EXPERIMENTAL RESULTS ON THE TEST SET OF THE MUSTARD DATASET.

Modality	Method	Speaker-Dependent			Speaker-Independent		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
T	MIARN	64.7	64.0	63.9	60.4	55.2	54.0
	BERT	67.5	66.9	66.8	58.2	56.7	57.0
	GPT-4o (CoT)	59.2	52.5	55.8	-	-	-
	LLaMA-3-8B (CoT)	46.2	43.7	44.9	-	-	-
T,V,A	EF-Concat	71.2	70.8	70.8	64.3	63.1	63.3
	IAIE	72.1	71.6	72.0	66.0	65.5	65.6
	IWAN	75.2	74.6	74.5	71.9	71.3	70.0
	CAAF	76.9	76.2	76.1	73.7	72.7	71.4
	IMRL& IMEIL	76.1	76.0	75.8	72.0	70.4	71.2
T,V,A	Ours	**78.7**	**78.1**	**77.9**	**74.2**	**73.4**	**73.1**

*T, V, and A mean textual, visual, and acoustic, respectively.

pooling along the sequence dimension to obtain the final multimodal inconsistency feature:

$$H_m = \sum_{i=1}^{n}(g_{tv} \odot H_{tv} + g_{ta} \odot H_{ta} + H_{tva}) \tag{11}$$

C. Knowledge-based Contextual Inconsistency Modeling

Sarcasm is characterized by high cultural relevance and context dependence. Addressing sarcasm's reliance on cultural context and logical subtleties, we introduce a framework coupling LLM-based knowledge retrieval with heterogeneous graph reasoning to capture deep semantic incongruities.

Daily conversations teem with proper nouns and cultural entities whose meanings often elude traditional models lacking background knowledge. To address this, we leverage ChatGPT's open-domain capabilities as an external knowledge source. Given a context-utterance sequence $\mathrm{U} = \{u_1, u_2, ..., u_t\}$, we prompt ChatGPT (Fig. 3) to extract entities and candidate concepts. For sentence $u_i \in \mathbb{R}^d$, entities are represented as $E_i \in \mathbb{R}^{m\times d}$ and concepts as $C_i \in \mathbb{R}^{m\times k\times d}$, with m and k denoting maximum entities and concepts per entity, respectively. Unlike prior work [19] that focuses solely on concept-entity attention, we emphasize the relevance of concepts to the specific sentence context, as entities may require different conceptual interpretations across contexts. We estimate concept importance by calculating the similarity between the concept C_i and sentence feature u_i. This contextual importance enhances the original entity features:

$$E_i' = E_i + \sum_{j=1}^{k} \mathrm{Softmax}(u_t \cdot C_i^j) \cdot C_i^j \tag{12}$$

To better capture long-range dependencies and the complex structural information of sarcasm within long contexts, we construct a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$ containing sentence nodes and entity nodes. The sentence nodes are denoted as $\mathrm{U}' = \{u_1', u_2', ..., u_t'\}$, and the enhanced entity nodes as $E' = \{e_1', e_2', ..., e_m'\}$. The graph contains two types of meta-paths to explicitly characterize the sources of inconsistency: Entity-Sentence-Entity (ESE), which considers semantic inconsistencies between different entities within a sentence, and Sentence-Entity-Sentence (SES), which reflects semantic inconsistencies between sentences containing the same entity. To facilitate information interaction between nodes and capture contextual inconsistencies, we introduce the node-level attention mechanism from Heterogeneous Graph Attention Networks [20]. We calculate sentence attention α and entity attention β for sentence nodes with SES meta-path structures and entity nodes with ESE meta-path structures, respectively. Furthermore, we apply a multi-head paradigm similar to multi-head attention to capture diverse features from different relations, as shown below:

$$\alpha_{ij}^k = \frac{exp(\eta(W_u^T \cdot [u_i'; u_j']))}{\sum_{l\in N_{u_i}} exp(\eta(W_u^T \cdot [u_i'; u_l']))} \tag{13}$$

$$\beta_{ij}^k = \frac{exp(\eta(W_e^T \cdot [e_i'; e_j']))}{\sum_{l\in N_{e_i}} exp(\eta(W_e^T \cdot [e_i'; e_l']))} \tag{14}$$

where u_i' and u_j' represent nodes with SES path, and e_i' and e_j' represent nodes with ESE path. W_u and W_e are learnable parameters, η is the nonlinear activation function LeakyReLU [21], and k denotes the number of attention heads. Finally, the output features are obtained via the following formulas:

$$u_i'' = \Big\Vert_{k=1}^{K} \sigma\Big(\sum_{j\in N(u_i')} \alpha_{ij}^k u_j'\Big) \tag{15}$$

$$e_i'' = \Big\Vert_{k=1}^{K} \sigma\Big(\sum_{j\in N(e_i')} \beta_{ij}^k e_j'\Big) \tag{16}$$

where u_i'' represents the feature representation containing sentence semantic inconsistency information learned by u_j'' from the SES meta-path, and e_i'' represents the feature representation containing entity semantic inconsistency information learned by e_j'' from the ESE meta-path. To automatically learn the differences and complementarity of the representations from the two meta-paths and achieve precise heterogeneous fusion, we introduce a co-attention mechanism [22]. First, we calculate the affinity matrix between u_i'' and e_i'':

$$A = tanh(U'' W_{ue} E''^T) \tag{17}$$

where $W_{ue} \in \mathbb{R}^{d \times d}$ is a learnable parameter, and $A \in \mathbb{R}^{t \times m}$ describes the bidirectional correlation between sentence nodes and entity nodes. Subsequently, utilizing the affinity matrix as an explicit interaction signal, we generate collaborative representations for sentences and entities, respectively:

$$H_u = tanh(U''W_1 + A(E''W_2)) \tag{18}$$

$$H_e = tanh(E''W_2 + A^T(U''W_1)) \tag{19}$$

$$\epsilon_u = softmax(H_u W_{hu}) \tag{20}$$

$$\epsilon_e = softmax(H_e W_{he}) \tag{21}$$

where W_1, W_2, W_{hu}, W_{he} are learnable parameters, and ϵ_u, ϵ_e represent the attention weights for sentence features and entity features, serving as the importance of the meta-paths. Based on these weights, we perform weighted average pooling on the nodes and concatenate them along the feature dimension to form the contextual semantic inconsistency feature M. Finally, we integrate the multimodal inconsistency feature H_m, the contextual inconsistency feature M, and utterance-level information features via concatenation, and output the prediction through a Multi-Layer Perceptron (MLP):

$$\hat{y} = softmax(W[H_m; M; V_u; T_u; A_u] + b) \tag{22}$$

where W is a learnable parameter and b is the bias.

IV. Experiments

A. Dataset

To evaluate the effectiveness of the proposed method, we conducted extensive experiments on the MUStARD dataset [4]. MUStARD is a benchmark dataset for multimodal sarcasm detection tasks, consisting of 690 conversation snippets collected from popular TV shows, including *Friends*, *The Big Bang Theory*, *The Golden Girls*, and *Sarcasm Anonymous*. The dataset exhibits a balanced class distribution, containing 345 sarcastic utterances and 345 non-sarcastic utterances. Each sample is a trimodal instance comprising visual, acoustic, and textual forms, providing rich information for capturing multimodal inconsistencies.

B. Implementation Details

Our framework is implemented in PyTorch. The input dimensions are determined by pre-trained backbones: textual dimension d_t is 768, visual dimension d_v is 2048, and acoustic dimensions d_a and d_{au} are 88 and 1000, respectively. The hidden layer dimension is set to 600. We set $L = 4$, $K = 2$, and the number of attention heads to 12. We employ Adam optimizer with learning rate 1×10^{-5} and dropout rate 0.3 to prevent overfitting.

C. Baselines

We evaluate the proposed method against baselines on MUStARD in Speaker-Dependent and Speaker-Independent settings. As shown in Table I, baselines fall into two groups: unimodal text methods and multimodal methods using text, audio, and vision. Unimodal baselines include MIARN [8], BERT [18], GPT-4o [23], and LLaMA-3-8B [24]. Multimodal baselines include EF-Concat [4], IAIE [12], IWAN [2], CAAF [3], and our reproduced IMRL & IMEIL [13]. For a fair and rigorous comparison, we report the optimal experimental results directly from the related literature for most baselines [2]–[4], [12], [25]. The only exception is IMRL & IMEIL [13]. We reproduced its results on the MUStARD dataset by utilizing the exact same feature extractors and data partitioning settings as our proposed method to ensure strict alignment.

TABLE II
Ablation study on modalities of our model and corresponding experimental results.

Setting	Modality	Precision	Recall	F1-Score
Speaker-Dependent	T, A	72.3	72.2	72.2
	T, V	72.9	72.3	72.8
	T, V, A	**78.7**	**78.1**	**77.9**
Speaker-Independent	T, A	68.1	67.5	67.3
	T, V	68.5	68.3	68.3
	T, V, A	**74.2**	**73.1**	**73.1**

TABLE III
Ablation Study Results of Different Modules.

Setting	Method	Precision	Recall	F1-Score
Speaker-Dependent	Ours	**78.7**	**78.1**	**77.9**
	w/o MI	77.0	75.6	76.3
	w/o CI	76.2	75.4	75.8
	w/o MI & CI	73.0	72.4	72.4
Speaker-Independent	Ours	**73.2**	**73.1**	**73.1**
	w/o MI	72.1	71.3	71.7
	w/o CI	71.7	70.9	71.3
	w/o MI & CI	70.6	69.6	67.8

*MI: Multimodal Inconsistency Module; CI: Contextual Inconsistency Module.

D. Overall Results

We validate our model under **Speaker-Dependent (S-D)** and **Speaker-Independent (S-I)** settings. Following the experimental settings of previous works, for S-D, we employ 5-fold cross-validation with random data partitioning. Each fold uses one subset for testing and four for training, with results averaged across all folds. For S-I, we reserve all *Friends* utterances as the test set and train on remaining data, ensuring complete speaker separation to eliminate identity bias. We evaluate using Precision, Recall, and F1-score.

Table I compares our proposed framework against baseline models on the MUStARD dataset. Baselines include unimodal text-based methods and multimodal approaches. Results demonstrate that our method achieves consistent improvements across both S-D and S-I settings, validating its effectiveness in multimodal sarcasm detection.

In the S-D setting, our model achieves an F1-score of 77.9. Comparing unimodal and multimodal approaches reveals the essential role of acoustic and visual modalities. The transition from text-only BERT to multimodal EF-Concat improves F1-score by 4.0 points, confirming that non-verbal cues are indispensable for sarcasm detection. However, simple feature concatenation proves insufficient, our method improves F1-score by 7.1 points over EF-Concat. This gain underscores

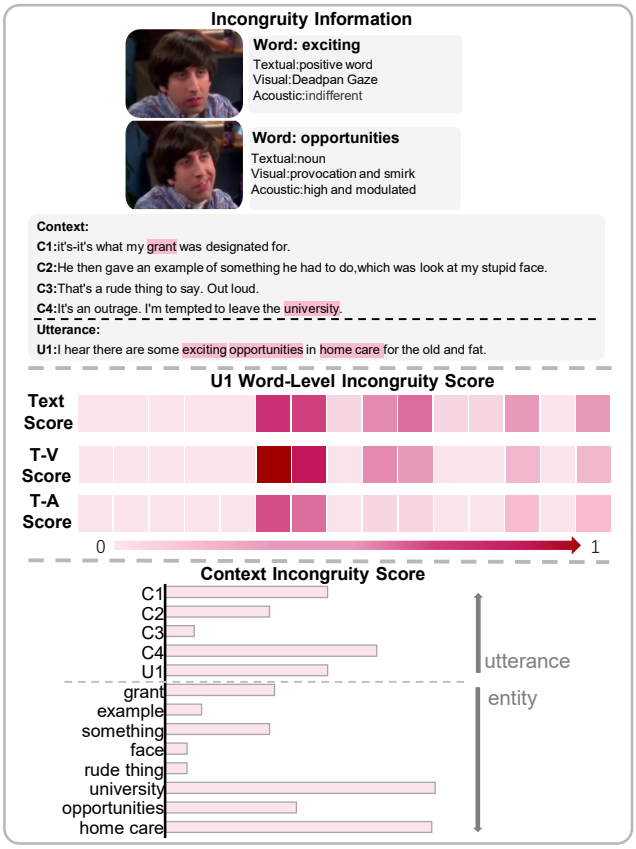

Fig. 4. Visualization of weight calculated by our model.

the importance of explicitly modeling cross-modal inconsistencies rather than merely combining features. Furthermore, our method outperforms the previous best multimodal baseline CAAF by 1.8% in F1-score, primarily attributed to our Multimodal Inconsistency Modeling module. By introducing inconsistency probes with diagonal attention constraints, our approach captures subtle token-level cross-modal conflicts more precisely than methods focusing solely on sentiment-bearing words.

Notably, LLM-based methods exhibit poor performance despite their powerful language understanding capabilities. LLaMA-3-8B achieves only 44.9 in F1-score, substantially underperforming even the early MIARN model which achieves 63.9, while GPT-4o with chain-of-thought prompting reaches 55.8 but still lags far behind multimodal approaches. This contrast reveals a fundamental limitation: text-only reasoning, regardless of model scale, cannot capture the cross-modal contradictions that are central to multimodal sarcasm. Our results validate that sarcasm detection inherently requires cross-modal perception rather than purely linguistic analysis.

The S-I setting poses a significantly greater challenge, as all test utterances come from unseen speakers in the *Friends* series. This speaker shift causes universal performance degradation: IWAN drops 4.5 points, CAAF drops 4.7 points, and IMRL&IMEIL drops 4.6 points in F1-score. Our method experiences a similar drop of 4.8 points, yet achieves the highest absolute performance with an F1-score of 73.1. This superior generalization stems from our Knowledge-based Contextual Semantic Inconsistency Mining module. Specifically, we leverage ChatGPT to enrich utterances with external commonsense knowledge and construct Sentence-Entity heterogeneous graphs to model deep semantic contradictions in context. By reasoning through logical inconsistencies rather than relying on speaker-specific cues such as habitual vocal tones or facial expressions, our approach maintains robust performance across different speakers. This demonstrates that knowledge-enhanced semantic reasoning provides more transferable representations for sarcasm detection than surface-level multimodal patterns.

E. Comparison of Input Modalities

To investigate the contribution of each non-verbal modality, we conducted experiments with different combinations of modalities. As shown in Table II, the full trimodal framework consistently achieves the best performance under both settings. Notably, removing either visual or acoustic information leads to a significant degradation in performance. For instance, in the S-D setting, the F1-score drops by 5.7% without visual cues and by 5.1% without acoustic cues. This decline confirms that both facial expressions and vocal intonations are indispensable for capturing the subtle inconsistencies necessary for sarcasm detection. Furthermore, the combination of text and video slightly outperforms the combination of text and audio, suggesting that in this dataset, visual cues may provide more distinct indications of sarcasm than vocal variations.

F. Ablation Study

To explore the contribution of each component within our proposed framework, we conducted an ablation study by separately removing the Multimodal Inconsistency Modeling module (MI) and the Knowledge-based Contextual Inconsistency module (CI). The results are presented in Table III. Removing the CI module resulted in a significant performance drop, with F1-scores decreasing by 2.1% and 1.8% in the two experimental settings, respectively. This indicates that capturing deep semantic discrepancies and cultural context via knowledge-enhanced heterogeneous graphs is crucial for the model's generalization capability, especially when specific speaker behavioral cues are unavailable. Similarly, removing the MI module also led to a decline in performance, with F1-scores dropping by 1.6% and 1.4% in the two settings, respectively. This suggests that the inconsistency probes play a complementary role by capturing fine-grained conflicts between textual and non-verbal cues, which is essential for detecting subtle sarcasm. Finally, the substantial drop in performance when both modules are removed demonstrates that these two components function synergistically, covering different aspects of sarcastic expression cues.

G. Case Study

To intuitively understand our model, we visualized the attention weights for the sarcastic utterance, context, and entities. As shown in Fig 4, for the word 'exciting', the text conveys positive emotion, whereas the speaker's vacant eyes

and deadpan gaze, paired with a flat tone, reveal his true intent. Our model successfully captured this inter-modal inconsistency and paid heightened attention to the information from all three modalities for this word. The sarcasm is even more pronounced for the word 'opportunities', where a provocative expression and high-pitched tone directly convey the sarcastic intent. Simultaneously, our model also captured inconsistency cues within the context, focusing more on entities like 'grant', 'university', 'opportunities', and 'home care', as well as sentences C1, C4, and U1. By leveraging external knowledge obtained from LLMs, the model identified the conflict between the high educational status associated with 'university' and the service industry nature of 'home care.' Combined with the sentence labeling the latter as an 'exciting opportunity', the model inferred the absurdity of the suggestion, exposing the speaker's intent to mock the other person's desire to leave the university. In summary, our model is capable of capturing not only inconsistencies between different modalities of words but also logical conflicts within the context.

V. Conclusion

This paper proposes a framework for multimodal sarcasm detection, addressing the challenges of fine-grained conflict detection and the scarcity of background knowledge. Our method introduces a Token-level Multimodal Inconsistency Modeling module, utilizing learnable probes to capture instantaneous cross-modal contradictions. Furthermore, we design a Knowledge-based Contextual Semantic Inconsistency Mining module, which leverages Large Language Models and heterogeneous graphs to uncover deep logical breaks within the dialogue context. Extensive experiments on the MUStARD dataset demonstrate that our framework achieves comparable performance. Ablation studies confirm the efficacy of our modules in integrating multimodal cues with deep semantic logic. Future work will explore incorporating speaker personality traits to further enhance interpretability and accuracy.

VI. Acknowledge

The authors would like to thank all the anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177, 62006167 and 62376181), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Nos. 24KJB520036 and 25KJA521001), Key R&D Plan of Jiangsu Province (BE2021048), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, "A multimodal corpus for emotion recognition in sarcasm," *arXiv:2206.02119*, 2022.

[2] Y. Wu, Y. Zhao, X. Lu, B. Qin, Y. Wu, J. Sheng, and J. Li, "Modeling incongruity between modalities for multimodal sarcasm detection," *IEEE MultiMedia*, vol. 28, no. 2, pp. 86–95, 2021.

[3] H. Xue, L. Xu, Y. Tong, R. Li, J. Lin, and D. Jiang, "Breakthrough from nuance and inconsistency: Enhancing multimodal sarcasm detection with context-aware self-attention fusion and word weight calculation," in *Proc. LREC-COLING*, 2024, pp. 2493–2503.

[4] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an obviously perfect paper)," in *Proc. ACL*, 2019, pp. 4619–4629.

[5] T. Veale and Y. Hao, "Detecting ironic intent in creative comparisons," in *Proc. ECAI*, 2010, pp. 765–770.

[6] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proc. EMNLP*, 2013, pp. 704–714.

[7] C. Van Hee, "Can machines sense irony? exploring automatic irony detection on social media," Ph.D. dissertation, Ghent University, 2017.

[8] Y. Tay, A. T. Luu, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," in *Proc. ACL*, 2018, pp. 1010–1020.

[9] C. Lou, B. Liang, L. Gui, Y. He, Y. Dang, and R. Xu, "Affective dependency graph for sarcasm detection," in *Proc. SIGIR*, 2021, pp. 1844–1849.

[10] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," in *Proc. COLING*, 2016, pp. 1601–1612.

[11] S. Attardo, J. Eisterhold, J. Hay, and I. Poggi, "Multimodal markers of irony and sarcasm," *Humor: Int. J. Humor Res.*, vol. 16, pp. 243–260, 2003.

[12] D. S. Chauhan, D. S R, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in *Proc. ACL*, 2020, pp. 4351–4360.

[13] D. Raghuvanshi, X. Gao, Z. Li, S. Bansal, M. Coler, N. Kumar, and S. Nayak, "Intra-modal relation and emotional incongruity learning using graph attention networks for multimodal sarcasm detection," in *Proc. ICASSP*, 2025, pp. 1–5.

[14] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. ICML*, 2023, pp. 19 730–19 742.

[15] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2017, pp. 4278–4284.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[19] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *Proc. ACM Multimedia*, 2019, pp. 1942–1951.

[20] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *Proc. WWW*, 2019, pp. 2022–2032.

[21] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv:1505.00853*, 2015.

[22] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. NeurIPS*, 2016.

[23] OpenAI, "Gpt-4 technical report," *arXiv:2303.08774*, 2023.

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.

[25] J. Lee, W. Fong, A. Le, S. Shah, K. Han, and K. Zhu, "Pragmatic metacognitive prompting improves LLM performance on sarcasm detection," in *Proceedings of the 1st Workshop on Computational Humor (CHum)*, C. F. Hempelmann, J. Rayz, T. Dong, and T. Miller, Eds. Online: Association for Computational Linguistics, Jan. 2025, pp. 63–70. [Online]. Available: https://aclanthology.org/2025.chum-1.7/