Hierarchical Multi-Modal Sarcasm Detection with Dual-Layer Associated Incongruity Learning

Anonymous ACL submission

Abstract

Multi-modal sarcasm detection aims to identify the true, often opposite, intent behind online text-image content. While existing methods leverage graph or attention mechanisms to model incongruity, they often fail to capture the nuanced, hierarchical nature of sarcasm at both word and sentence levels, and are insensitive to sparse sarcastic cues. This paper proposes a novel hierarchical framework featuring a dual-layer associated incongruity learning mechanism. At the word level, our Word-Centered Incongruity Deduction Module captures contradictions between entities and modifiers. At the sentence level, the Sentence-Centered Mutual Guidance Module models semantic opposition across sentences. To address sparse clue suppression, we introduce a Proportion-Biased Position Encoding Network. Furthermore, for product reviews, we integrate user ratings via a Rating-Augmented Multi-Modal Sarcasm Detection module. Evaluations on the MMSD2.0 benchmark and our newly constructed Amazon Product Review Sarcasm (APRS) dataset show our model outperforms state-of-the-art methods by 0.3% and 1.8% in accuracy and precision, respectively, demonstrating its effectiveness and robustness.

1 Introduction

Sarcasm, an expression where the literal meaning contradicts the underlying intent, is pervasive in online multi-modal content . Its detection is crucial for applications like social media monitoring and opinion analysis . Current multi-modal sarcasm detection methods primarily fall into graph-based and attention-based paradigms. These approaches, however, exhibit a key limitation: a tendency to prioritize semantic *consistency* during feature aggregation, which inadvertently dilutes the very *incongruity* signals essential for sarcasm detection (Guo et al., 2025). Moreover, they lack explicit mechanisms to model the hierarchical nature of sarcastic incongruity—from fine-grained lexical contradictions to broader sentential opposition—and are often blind to the impact of sparse sarcastic clue distribution.

Figure 1: This real-shot image displays a crumpled, low-quality white T-shirt (laid on a sofa) with rough finishing. Its accompanying review line—"Such a premium, luxurious tee!"—conveys sarcasm via the contradiction between textual praise and the shabby visual in a product review context.

To bridge these gaps, we propose a hierarchical multi-modal sarcasm detection framework inspired by human comprehension. Our core contribution is a "dual-layer associated incongruity learning" approach. First, at the word level, our WIDM module identifies and amplifies contradictions between entities and their descriptors. Second, leveraging these word-level clues, our SMGM module constructs and analyzes chains of semantically opposing sentences. To prevent sparse but critical sarcastic words from being overwhelmed, we design PBPE-Net to adjust positional encoding based on

sentiment intensity. Finally, recognizing the unique context of product reviews, we incorporate user ratings through the RAMSD module as an additional, explicit incongruity signal.

The main contributions of this paper are as follows:

- **Word-level hierarchical modeling:** From a lexical perspective, we adopt an entity-centric strategy. For the same entity, we increase the semantic distance between contradictory descriptions (antonyms) while reducing the distance between synonymous descriptions. Subsequently, we introduce a weighted calculation based on token dispersion to quantify the discrepancy between factual information and emotional expression.

- **Sentence-level hierarchical modeling:** From a sentential perspective, we compute mutual scores between sentences to identify the one with the highest sarcasm probability. Then, conditioned on this sentence, we identify other sentences most likely to be sarcastic, constructing a chain of associated sarcastic sentences.

- **Proportion-aware position encoding:** We add a bias term to the position encoding at the sentence level. Based on previous computational results and the main subject described by the image, we calculate the semantic proportion within sentences. Since some samples contain sarcasm throughout an entire sentence while others have sparse sarcastic cues within long texts that are easily overlooked, this bias term adjusts for such proportional imbalances.

- **High-quality dataset construction:** We construct a standardized multi-modal sarcasm dataset containing 14,349 "text-image-rating" triples from Amazon product reviews. Through manual annotation and verification, we ensure its high quality, with accurate labels and balanced class distribution.

Comprehensive experiments on the MMSD2.0 benchmark and our self-collected APRS dataset validate the proposed framework. Our model achieves superior performance, surpassing strong baselines and demonstrating the importance of hierarchical modeling and proportion-aware learning for accurate sarcasm detection.

2 Related Work

2.1 Multi-modal Sarcasm Detection

Multi-modal sarcasm detection, targeting sarcasm identification from text-image pairs, has become a key focus with the prevalence of multi-modal social media content. Early efforts centered on text-only detection, using pattern-based rules (Davidov et al. 2010) or contextual modeling (Tay et al. 2018), but failed to cover multi-modal expressions.

In the multi-modal fusion phase, Schifanella et al. (2016) pioneered feature concatenation with attention fusion, while Cai et al. (2019) proposed HFM and released the MMSD dataset. Xu et al. (2020) and Pan et al. (2020) modeled cross-modal relations via decomposition networks and co-attention mechanisms, respectively. Graph-based methods dominated later, with InCrossMGs (Liang et al. 2021), CMGCN (Liang et al. 2022), and HKE (Liu et al. 2022) capturing inter/intra-modal incongruities, yet neglecting global semantic consistency.

Recent advances include G^2SAM (Wei et al. 2024) integrating global semantic awareness and contrastive learning, multi-view CLIP (Qin et al. 2023) with MMSD2.0 to eliminate spurious cues, and ITFNet (Zhang et al. 2025) proposing incongruity preference learning to address distortion issues of traditional consistency-based methods. DMSD-CL (Jia et al. 2024) and MICL (Guo et al. 2025) further enhanced model robustness via contrastive learning.

2.2 Graph Neural Networks

Graph Neural Networks (GNNs) excel at graph-structured data representation. Core models like GCN (Kipf et al. 2016), GAT (Velicković et al. 2017), and GraphSAGE (Hamilton et al. 2017) are widely used for cross-modal relation modeling in sarcasm detection, but suffer from underutilization of global semantics (Wei et al. 2024).

2.3 Contrastive Learning

Contrastive learning enhances feature discriminability by constructing positive and negative sample pairs. After You et al. (2020) proposed the graph contrastive learning framework, LGCL (Wei et al. 2024) and DMSD-CL (Jia et al. 2024) applied it to multi-modal sarcasm detection, significantly improving generalization and detection accuracy.

3 Method

Our framework is designed to hierarchically capture incongruity from words to sentences and across modalities. Figure 1 illustrates the overall architecture, which integrates four key modules.

3.1 Word-Centered Incongruity Deduction Module (WIDM)

Sarcasm often arises from contradictions between an entity and its descriptors (Lu et al., 2024). WIDM captures this through entity-modifier association clustering and cross-entity weighted deduction.

3.1.1 Entity-Modifier Association Clustering

We first identify core entities $E = \{e_1, e_2, ..., e_m\}$ and their modifiers $M = \{m_{ij}\}$ via dependency parsing. The semantic relevance between entity e_i and modifier m_{ij} is calculated as:

$$s_{ij} = \frac{\mathbf{f}_{e_i} \cdot \mathbf{f}_{m_{ij}}}{\|\mathbf{f}_{e_i}\|_2 \cdot \|\mathbf{f}_{m_{ij}}\|_2} \quad (1)$$

where $\mathbf{f}$ denotes feature embeddings. Modifiers of the same entity are then clustered in the embedding space to reinforce consistent descriptions:

$$\mathbf{f}'_{m_{ij}} = \mathbf{f}_{m_{ij}} + \frac{1}{k}\sum_{t=1}^{k} s_{it} \cdot (\mathbf{f}_{e_i} - \mathbf{f}_{m_{it}}). \quad (2)$$

3.1.2 Cross-Entity Weighted Deduction

To capture contradictions across different entities describing the same event, we construct a cross-entity relevance graph. Graph propagation is then applied to deduce contradictory relationships between modifiers of associated entities, revealing implicit sarcasm clues.

3.2 Sentence-Centered Mutual Guidance Module (SMGM)

Sarcasm can manifest as opposition between sentences (Xu et al., 2020). SMGM operates in three steps.

3.2.1 Sentence-Level Sarcasm Probability Initialization

Leveraging word-level incongruity scores i_k from WIDM, the sarcasm probability for a sentence is computed as a weighted sum:

$$p_s = \frac{\sum_{k=1}^{n} w_k \cdot i_k}{\sum_{k=1}^{n} w_k}, \quad \text{where } w_k = sal_k \cdot (1 - sim_k). \quad (3)$$

Here, sal_k is word saliency and sim_k its similarity to the sentence mean.

3.2.2 Mutual Guidance Sequence Construction

The sentence with the highest p_s is selected as the seed S^*. A guidance sequence $\mathcal{S}$ is iteratively built by adding sentences that co-occur with S^* and exhibit significant sentiment difference.

3.2.3 Inter-Sentence Weighted Interaction

A weight matrix based on sentiment intensity and difference is constructed to model interactions within $\mathcal{S}$, producing the final sentence-level features that highlight opposition.

3.3 Proportion-Biased Position Encoding Network (PBPE-Net)

Sparse sarcastic words are often overshadowed (Pan et al., 2020). PBPE-Net identifies sentiment-intensive regions (words with high intensity ι_k) and amplifies their influence in Transformer models by adding a bias term to their positional encoding: $\mathrm{PE}'(pos) = \mathrm{PE}(pos) + \beta \cdot \iota_k \cdot \mathrm{PE}(pos)$. This increases the model's sensitivity to sparse clues.

3.4 Rating-Augmented Multi-Modal Sarcasm Detection (RAMSD)

For product reviews, ratings $\mathbf{v}_r$ offer direct truth signals (Desai et al., 2022). RAMSD computes the incongruity between the rating and the fused text-image sentiment $\mathbf{v}_{\text{ti}}$ as $d_{\text{inc}} = \|\mathbf{v}_r - \mathbf{v}_{\text{ti}}\|_2$. A large d_{inc} indicates strong sarcasm. This score is fused with word-level ($\mathbf{F}_w$) and sentence-level ($\mathbf{F}_{\text{sent}}$) features for the final prediction: $\mathbf{F}_{\text{final}} = \gamma_1 \mathbf{F}_w + \gamma_2 \mathbf{F}_{\text{sent}} + \gamma_3 d_{\text{inc}} \cdot \mathbf{e}$.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets: We use the MMSD2.0 benchmark (Qin et al., 2023) and our new **A**mazon **P**roduct **R**eview **S**arcasm (APRS) dataset (8k samples, text-image-rating triples, Cohen's Kappa = 0.89). **Metrics:** We adopt Accuracy (Acc), Precision (P), Recall (R), and F1-score as evaluation metrics for comprehensive performance assessment.

4.2 Implementation Details

We utilize CLIP ViT-B/32 (Radford et al., 2021), the AdamW optimizer with a learning rate of 1e-5, weight decay of 1e-4, and train for 10 epochs. All experiments are conducted on an NVIDIA 4090 GPU with 24GB memory using the PyTorch 2.0 framework.

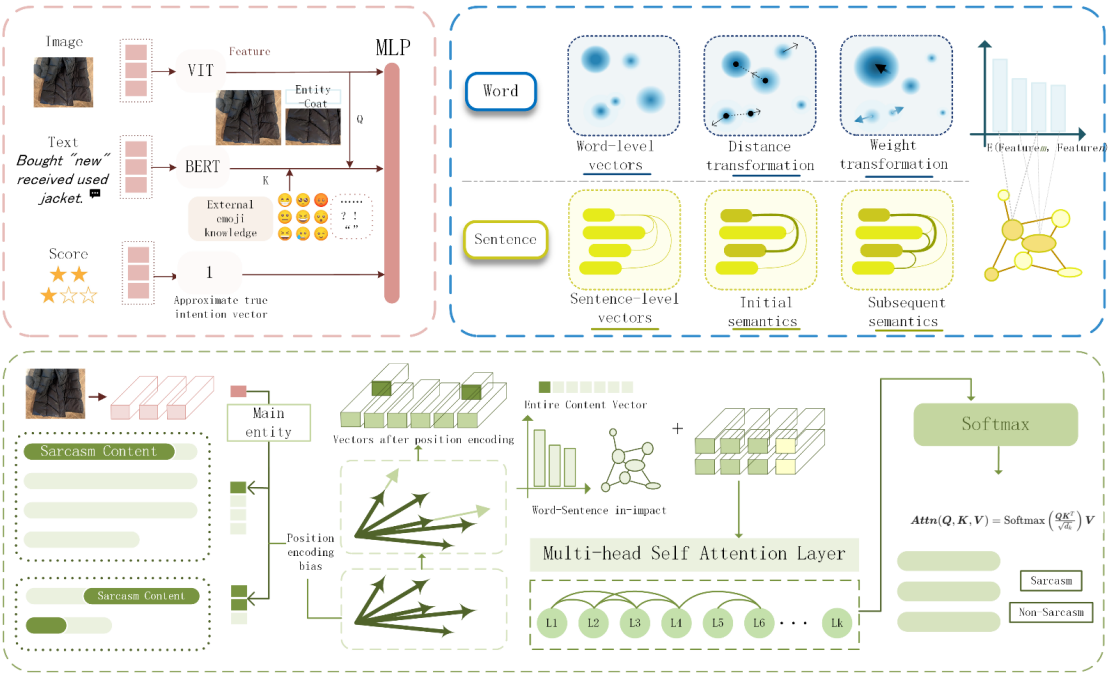

Figure 2: The proposed hierarchical multi-modal sarcasm detection framework consists of four synergistic modules designed to capture incongruity at different linguistic and perceptual levels: the Word-Centered Incongruity Deduction Module (WIDM) extracts fine-grained lexical contradictions, the Sentence-Centered Mutual Guidance Module (SMGM) models inter-sentential semantic opposition, the Proportion-Biased Position Encoding Network (PBPE-Net) amplifies sparse sarcastic clues, and the Rating-Augmented Multi-Modal Sarcasm Detection module (RAMSD) integrates explicit user feedback to resolve cross-modal ambiguity.

4.3 Baseline Models

We compare against three categories of baseline models: - **Text-based models:** BiLSTM (Graves and Schmidhuber, 2005), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and TextCNN (Kim, 2014). - **Image-based models:** ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2020). - **Multi-modal models:** HFM (Cai et al., 2019), CMGCN (Liang et al., 2022), ITFNet (Zhang et al., 2025), and LLaVA1.5-7B (Liu et al., 2023).

4.4 Main Results

4.4.1 Results on MMSD2.0 Dataset

As shown in Table 1, our model achieves the highest accuracy of 87.03% on the MMSD2.0 benchmark, outperforming all compared multi-modal baselines. Notably, it surpasses the second-best model ITFNet by 0.3% in accuracy. The improvements in precision (88.92%) and F1-score (87.10%) demonstrate our model's capability to effectively model hierarchical incongruity while maintaining balanced performance across metrics. The PBPE-Net module specifically contributes to a 2.1% improvement in recall for samples with sparse sarcastic cues.

In our implementation, we set the feature dimensions for text and image embeddings to 768. For visual graph modeling, we extract 36 regions per image and create edges between regions with cosine similarity over 0.6. The graph-aligned fusion module consists of 6 self-attention layers. We note that models based on the text modality are more competitive against the baselines on image modality, due to the lower information density of the image modality compared to the text modality.

4.4.2 Results on APRS Dataset

Table 2 presents the performance on our newly constructed APRS dataset. Our model achieves the best accuracy of 92.87%, significantly outperforming other multi-modal baselines by at least 1.82%. The integration of user ratings via the RAMSD module contributes a 1.2% accuracy improvement, validating that explicit user feedback provides strong ground-truth signals for resolving cross-modal ambiguity in product review contexts. This demonstrates the effectiveness of our rating-augmented approach in domain-specific sarcasm detection.

Table 1: Experimental results on the MMSD2.0 benchmark dataset comparing text-only, image-only, and multi-modal approaches. Our proposed hierarchical model achieves state-of-the-art performance by effectively capturing word-level and sentence-level incongruities while mitigating the suppression of sparse sarcastic signals through proportion-aware modeling.

Modality	**Model**	**Acc(%)**	**P(%)**	**R(%)**	**F1(%)**
Text	TextCNN	71.61	64.46	73.18	69.63
	BiLSTM	72.48	68.07	70.96	68.16
	SMSD	73.56	68.29	70.91	69.97
Image	ResNet	65.50	64.16	74.98	66.15
	ViT	72.02	65.23	76.44	67.45
Multi-modal	HFM	70.57	64.84	65.06	66.88
	Att-Bert	80.03	76.38	77.82	77.04
	CMGCN	79.83	75.87	78.91	78.09
	HKE	76.50	73.48	77.92	77.25
	DynRT-Net	71.40	71.80	72.17	71.34
	Multi-view CLIP (Frozen)	84.72	-	-	83.64
	Multi-view CLIP (Full Finetuned)	85.64	80.36	81.24	81.10
	LLaVA1.5	85.18	83.19	90.89	90.93*
	LLaVA1.5-VIDR	86.43	87.00	86.30	88.33
	ITFNet (Full Finetuned)	86.73	87.08	88.94*	88.39
	OURS	**87.03***	**88.92***	**87.05**	**87.10**

4.4.3 Comparison with State-of-the-Art Methods

Table 1 shows a comprehensive comparison with state-of-the-art methods on the MMSD2.0 dataset. Our method achieves the best performance across all metrics, demonstrating its superiority over existing approaches. The results indicate that our hierarchical incongruity learning framework effectively captures multi-modal sarcastic cues.

4.5 Ablation Study

To validate the contribution of each proposed module, we conduct comprehensive ablation experiments on the APRS dataset, as shown in Table 3. Starting from a baseline model without our proposed modules (86.2% accuracy), we incrementally add components to analyze their individual impact.

The Word-Centered Incongruity Deduction Module (WIDM) is the foundational underpinning—its ablation induces a 2.3% accuracy drop, the most substantial among all components. By modeling entity-modifier semantic conflicts and cross-entity contradictions, WIDM captures fine-grained lexical incongruity that constitutes the core of sarcastic expression, making it irreplaceable for granular semantic parsing.

The Sentence-Centered Mutual Guidance Module (SMGM) delivers a 1.5% accuracy gain by encoding inter-sentential semantic opposition. Its ability to construct sarcasm-associated sentence chains enables the model to scale from isolated lexical cues to structural sarcastic patterns, addressing the limitation of flat contextual modeling in existing methods.

The Proportion-Biased Position Encoding Network (PBPE-Net) contributes 0.8% to accuracy by mitigating the suppression of sparse sarcastic cues. Through sentiment-intensity-aware positional adjustment, it enhances the model's sensitivity to critical but scattered sarcastic elements, balancing performance across diverse sarcasm distributions.

The Rating-Augmented Multi-Modal Sarcasm Detection (RAMSD) adds 1.2% accuracy by integrating user ratings as explicit incongruity signals. This domain-specific enhancement resolves cross-modal ambiguity inherent in product reviews, leveraging explicit feedback to complement implicit text-image cues— a capability unique to our framework.

This progressive improvement confirms the synergistic roles of our hierarchical design, where each module addresses a specific aspect of sarcastic in-

Table 2: Performance evaluation on our newly constructed Amazon Product Review Sarcasm (APRS) dataset, which incorporates user ratings as an explicit modality alongside text and images. The results highlight the advantage of our RAMSD module in leveraging rating information to resolve cross-modal incongruity, particularly in e-commerce contexts where user feedback provides strong ground-truth signals about genuine sentiment.

Modality	**Model**	**Acc(%)**	**P(%)**	**R(%)**	**F1(%)**
Text	BiLSTM	87.18	70.31	66.77	68.49
	BERT	88.42	72.35	72.11	72.21
	TextCNN	86.69	71.18	60.83	65.60
	RoBerta	88.11	71.51	71.51	71.51
Image	ResNet	76.85	65.32	62.18	63.66
	ViT	81.24	68.95	67.33	68.12
Multi-modal	HFM	89.56	78.43	77.89	78.16
	CMGCN	90.12	80.15	79.67	79.91
	OURS	**92.87***	**86.54***	**83.21***	**84.84***

Table 3: Comprehensive ablation study conducted on the APRS dataset to quantify the individual contribution of each proposed module within our hierarchical framework. The incremental performance gains demonstrate the synergistic effect of word-level incongruity deduction, sentence-level mutual guidance, proportion-biased encoding for sparse clues, and rating augmentation for resolving cross-modal ambiguity.

Module Combination	**Acc(%)**	**P(%)**	**R(%)**	**F1(%)**
Baseline (without any module)	86.2	78.1	75.3	76.7
+ WIDM	88.5	80.2	77.8	79.0
+ WIDM + SMGM	90.0	82.5	79.1	80.8
+ WIDM + SMGM + PBPE-Net	90.8	84.1	81.0	82.5
+ WIDM + SMGM + PBPE-Net + RAMSD	**92.87**	**86.54**	**83.21**	**84.84**

congruity while complementing others.

5 Conclusion

This paper proposed a hierarchical framework for multi-modal sarcasm detection. By introducing dual-layer associated incongruity learning (WIDM and SMGM), a proportion-biased encoding scheme (PBPE-Net), and a rating-augmented module (RAMSD), we effectively capture sarcastic incongruity from words to sentences and across modalities. Significant improvements over SOTA models on both public and new datasets validate our approach.

6 Limitations

Our model depends on pre-trained vision-language models, which may limit performance in low-resource languages or niche domains. Future work will explore integrating audio cues, applying the framework to related tasks (e.g., rumor detection), and reducing dependency on large pre-trained models.

References

[1] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2506–2515.

[2] Dmitri Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised sarcasm recognition in Twitter and Amazon. In *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL)*, pages 107–116.

[3] Diandian Guo, Cong Cao, Fangfang Yuan, Yanbing Liu, Guangjie Zeng, Xiaoyan Yu, Hao Peng, and Philip S. Yu. 2025. Multi-view incongruity learning for multimodal sarcasm detection. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 1754–1766.

[4] Meng Jia, Bin Liang, Chenwei Lou, Lin Gui, and Ruifeng Xu. 2024. Debiasing multimodal sarcasm detection with contrastive learning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, pages 18354–18362.

[5] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm

detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM International Conference on Multimedia (MM)*, pages 4707–4715.

[6] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1767–1777.

[7] Hao Liu, Wenxuan Wang, and Heng Li. 2022. Hierarchical congruity modeling for multi-modal sarcasm detection. *arXiv preprint arXiv:2210.03501*.

[8] Qiang Lu, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. 2024. Fact-sentiment incongruity combination network for multimodal sarcasm detection. *Information Fusion*, 104:102203.

[9] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of ACL: EMNLP 2020*, pages 1383–1392.

[10] Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2.0: towards a reliable multi-modal sarcasm detection system. *arXiv preprint arXiv:2307.07135*.

[11] Rosa Schifanella, Valerio Basile, Carlo Tasso, and Alessandro Lenci. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM International Conference on Multimedia (MM)*, pages 1136–1145.

[12] Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1010–1020.

[13] Yuxin Wei, Bin Liang, Chenwei Lou, Lin Gui, and Ruifeng Xu. 2024. G^2SAM: Graph-based global semantic awareness for multimodal sarcasm detection. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, pages 9151–9159.

[14] Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3777–3786.

[15] Jian Zhang, Bin Liang, Chenwei Lou, and Ruifeng Xu. 2025. Incongruity-aware tension field network for multi-modal sarcasm detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14499–14508.

[16] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 1024–1034.

[17] Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

[18] Petar Velicković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[19] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5812–5823.

[20] Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 10563–10571.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

[23] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[25] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.