# GDCNET: GENERATIVE DISCREPANCY COMPARISON NETWORK FOR MULTIMODAL SARCASM DETECTION

*Shuguang Zhang, Junhong Lian, Guoxin Yu, Baoxun Xu, Xiang Ao*[*]

State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
Shenzhen Stock Exchange

## ABSTRACT

Multimodal sarcasm detection (MSD) aims to identify sarcasm within image–text pairs by modeling semantic incongruities across modalities. Existing methods often exploit cross-modal embedding misalignment to detect inconsistency, but struggle when visual and textual content are loosely related or semantically indirect. Recent approaches leverage large language models (LLMs) to generate sarcastic cues or explanations, but the diverse perspectives of the generated sarcastic text from LLMs may amplify noise. To remedy these issues, we propose the **G**enerative **D**iscrepancy **C**omparison **N**etwork (**GDCNet**), a simple yet effective framework that captures cross-modal conflicts by introducing descriptive and factually grounded image captions generated by LLMs as stable semantic anchors. Specifically, GDCNet first employs a multi-modal large language model to generate factual, image-consistent textual descriptions. Then, it computes semantic and sentiment-level discrepancies between the generated description and the original text, while also measuring the visual-textual consistency. These discrepancies are fused with visual and textual features via a gated module to adaptively balance modality contributions. Extensive experiments on MSD benchmarks validate GDCNet's substantial gains in accuracy and robustness, setting a new state-of-the-art on the MMSD2.0 benchmark. Our framework implementation and code are available at `https://anonymous.4open.science/r/GDCNet`.

***Index Terms*—** Multimodal sarcasm detection, discrepancy generation, multimodal representation enhancement

## 1. INTRODUCTION

Sarcasm is a linguistic phenomenon in which the surface meaning of an utterance diverges significantly from the speaker's intended communicative message. It is commonly employed for humor, critique, and subtle social commentary [1], making it a potent tool for communication. The surge of multimodal social-media content has subsequently driven the advancement of Multimodal Sarcasm Detection (MSD) as an emerging research frontier. Transitioning from text-only to multimodal contexts considerably complicates this task, as the interplay between images and text often generates irony that transcends the meaning conveyed by either modality in isolation.

MSD fundamentally relies on identifying the incongruity between images and text. Prior studies have tackled this challenge by aligning multimodal representations [2], leveraging techniques such as attention mechanisms [3], graph neural networks [4], external knowledge [5], and dynamic routing [6]. Despite these advances, existing methods [3, 4, 5, 6] still struggle with out-of-distribution generalization and often rely on superficial cues [7]. Their reliance on cross-modal embedding inconsistencies captures broad misalignments but often misses the subtle ironic cues crucial for accurate sarcasm detection, especially when image–text alignment is weak.

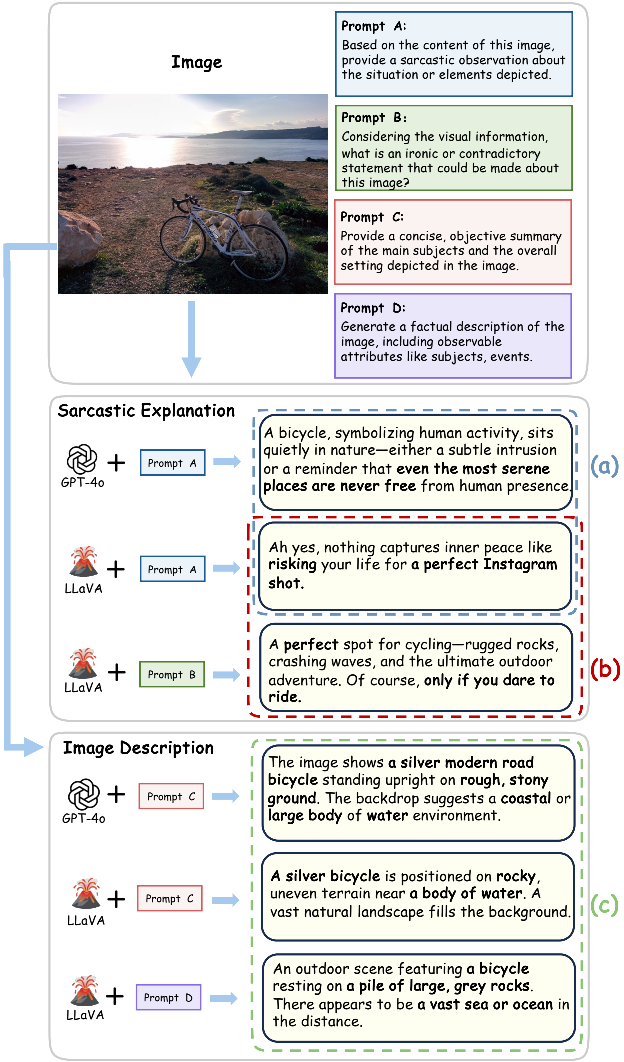
Recent successes in Large Language Models (LLMs) and their multimodal extensions (MLLMs) [8, 9] have introduced extensive world knowledge and cross-modal reasoning capabilities, providing renewed impetus for the MSD task. These approaches generally exploit prompts to guide MLLMs in generating sarcastic explanations or signals for contextual data augmentation [10]. Such methods help alleviate the limitations of traditional cross-modal embedding mismatches, but they often overlook a fundamental challenge in sarcasm detection: the inherent diversity of representation perspectives. As established in cognitive psychology [1], this diversity stems from variations in human cognition and interpretation. The issue becomes even more pronounced in the MSD task, because distinct visual elements can contribute to generating varied ironic texts through culturally-dependent interpretations [11]. An illustrative example is exhibited in Fig. 1. From Fig. 1 (a), we can observe that different MLLMs generate completely different sarcastic expressions given the same picture. Even for the same MLLM, different prompts can guide it to generate distinct irony texts (c.f. Fig. 1 (b)). Based on such observation, our motivation is to explore a more robust way to utilize the capability of MLLMs to underpin the MSD task, following the idea of taking MLLMs as cross-modal semantic connectors.

In this paper, we propose a framework named **G**enerative **D**iscrepancy **C**omparison **Net**work (GDCNet for short). Inspired by the success of MLLMs in image captioning [12], we apply the MLLM to generate descriptive and factually grounded image captions, which serve as a cross-modality semantic bridge within our framework. These descriptions not only preserve visual semantics but also provide consistent and reliable semantic anchors for effective comparison with the associated textual content, as illustrated in Fig. 1(c). Specifically, within GDCNet, we first employ an MLLM to produce objective image descriptions that serve as stable semantic references. Based on these descriptions and the corresponding original text, we introduce a **G**enerative **D**iscrepancy **R**epresentation **M**odule (GDRM), which captures discrepancies across semantic and emotional alignment between texts, as well as image-text consistency at the representation level. To further enhance sarcasm detection performance, a gated fusion module is employed to integrate these multi-dimensional discrepancy representations with the original visual and textual features, thereby improving both the robustness and accuracy of the final classification.

Our main contributions are summarized as follows:

- We present GDCNet, a novel framework for MSD that leverages factual-grounded generated image descriptions as the semantic anchor to simplify incongruity detection across visual and textual modalities.

[*] Corresponding author.



**Prompt A:** Based on the content of this image, provide a sarcastic observation about the situation or elements depicted.

**Prompt B:** Considering the visual information, what is an ironic or contradictory statement that could be made about this image?

**Prompt C:** Provide a concise, objective summary of the main subjects and the overall setting depicted in the image.

**Prompt D:** Generate a factual description of the image, including observable attributes like subjects, events.

A bicycle, symbolizing human activity, sits quietly in nature—either a subtle intrusion or a reminder that **even the most serene places are never free** from human presence.

Ah yes, nothing captures inner peace like **risking** your life for **a perfect Instagram shot.**

A **perfect** spot for cycling—rugged rocks, crashing waves, and the ultimate outdoor adventure. Of course, **only if you dare to ride.**

The image shows **a silver modern road bicycle** standing upright on **rough, stony ground**. The backdrop suggests a **coastal** or **large body** of **water** environment.

**A silver bicycle** is positioned on **rocky**, uneven terrain near **a body of water**. A vast natural landscape fills the background.

An outdoor scene featuring **a bicycle** resting on **a pile of large, grey rocks**. There appears to be **a vast sea or ocean** in the distance.

**Fig. 1**. LLM outputs on the same image: sarcastic explanations diverge with models and prompts, while factual descriptions remain stable, highlighting their potential as reliable semantic anchors.

- We propose a GDRM to extract key representations for MSD by comparing semantic and sentiment differences between generated image descriptions and the original text, as well as assessing image-text consistency.
- Extensive experiments on widely-used benchmark demonstrate that our GDCNet achieves significant improvements in detection accuracy and establishes a new state-of-the-art on the MMSD2.0 dataset.

## 2. METHODOLOGY

### 2.1. Problem Formulation

Given an image-text pair $(I, T)$, the multimodal sarcasm detection task aims to classify it as sarcastic or non-sarcastic. Formally, this is formulated as a binary classification problem where we seek to learn a mapping $f : (I, T) \mapsto y$, with $y \in 0, 1$. Here, $y = 1$ indicates sarcastic and $y = 0$ otherwise. The core challenge is to identify cross-modal incongruities, often manifesting as the subtle mismatches between the textual semantics of $T$ and the visual context of $I$ for an input pair $(I, T)$.

### 2.2. Cross-modal Feature Alignment

Given the $i$-th sample $(I_i, T_i)$, we utilize modality-specific encoders $E_v$ and $E_t$ to produce visual features $h_i^v \in \mathbb{R}^{d_v}$ and textual features $h_i^t \in \mathbb{R}^{d_t}$. To align image and text features across modalities, we project both $h_i^v$ and $h_i^t$ into a shared latent space through learnable linear layers, obtaining $z_i^v$ and $z_i^t$ with the same dimensionality $d_z$. We then adopt contrastive learning to align the projected features. The similarity score $s_{ij}$ between $i$-th image and $j$-th text is defined by cosine similarity:

$$s_{ij} = \frac{(z_i^v)^\top z_j^t}{\|z_i^v\|_2 \, \|z_j^t\|_2}, \quad i, j \in \{1, \dots, B\}, \tag{1}$$

where $B$ is the mini-batch size. The margin-based contrastive loss is then:

$$\mathcal{L}_{\text{cont}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{\substack{j=1 \\ j \neq i}}^{B} \max\big(0,\ m + s_{ij} - s_{ii}\big), \tag{2}$$

where $m > 0$ is a margin hyperparameter encouraging separation between positive and negative pairs. This objective increases similarity of matched pairs while pushing apart mismatched ones, thereby enhancing cross-modal alignment.

### 2.3. Generative Discrepancy Representation Module

The Generative Discrepancy Representation Module (GDRM) captures implicit conflicts between the original text $T$ and the image $I$. It first employs an MLLM such as LLaVA-NEXT [9] to generate a textual description $\hat{T}$ from the image, ensuring that the description faithfully reflects visual semantics. To avoid sarcasm-related biases, the MLLM is restricted to image-only input, excluding any multimodal contextual cues. As a result, $\hat{T}$ serves as an unbiased and context-independent representation of the image.

To quantify the inconsistency between the generated description $\hat{T}$ and the original text $T$, we compute three discrepancies: semantic discrepancy, sentiment discrepancy, and visual-textual fidelity.

The semantic discrepancy $d_{\text{sem}}$ measures the divergence in meaning between the original text and the generated description. We compute the cosine dissimilarity between the CLIP text embeddings of the original text $T$ and the generated description $\hat{T}$. The sentiment discrepancy $d_{\text{sen}}$ captures shifts in sentiment between the original text and the generated description. We use a RoBERTa-based sentiment classifier [13] to obtain their sentiment probability distributions and compute the discrepancy as the $L_1$ distance. In addition, we measure visual-textual fidelity $d_{\text{fidelity}}$, i.e., the alignment between the generated description and the image. This is quantified as the cosine similarity $\cos(z_I, z_{\hat{T}})$ between the CLIP image embedding $z_I$ and the CLIP text embedding of $z_{\hat{T}}$. A lower $d_{\text{fidelity}}$ value signifies greater deviation between the generated text and the visual content, indicating a lack of visual fidelity.

The three discrepancies are concatenated to form a discrepancy feature vector:

$$D = d_{\text{sem}} \oplus d_{\text{sen}} \oplus d_{\text{fidelity}}, \tag{3}$$

where $\oplus$ denotes vector concatenation. The vector $D$ is then further processed by a multilayer perceptron (MLP) to obtain the final discrepancy representation $F_D$.

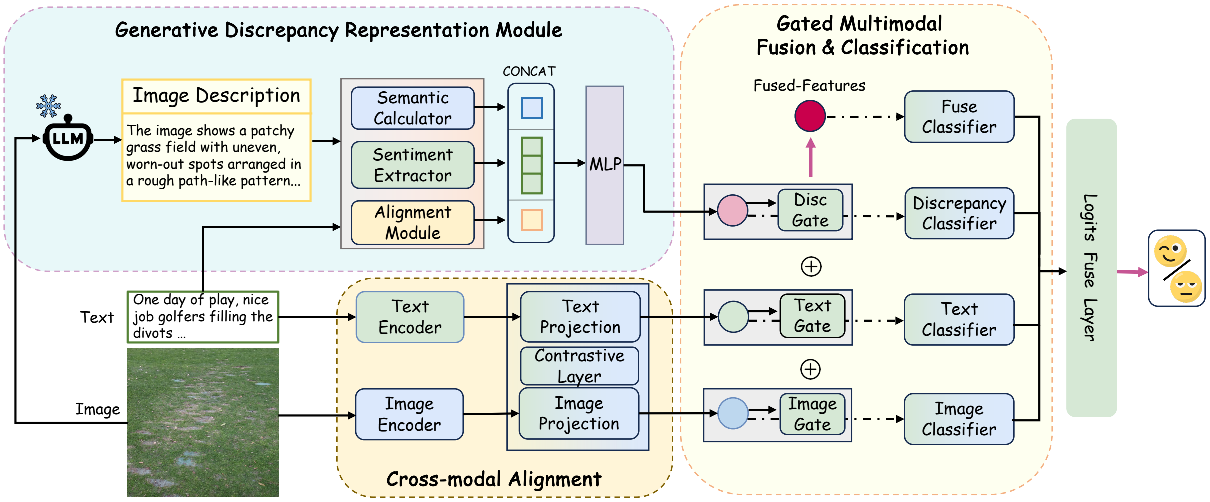


**Fig. 2**. The Architecture of GDCNet. In the Gated Multimodel Fusion & Classification module, the four circles in different colors represent discrepancy features, text features, image features, and fused features, respectively.

### 2.4. Gated Multimodal Fusion & Classification

To effectively integrate textual, visual, and discrepancy-based features, we employ the gated fusion mechanism. It assigns learnable importance weights to each modality, allowing the model to adaptively focus on the most informative features. Given feature vectors from the text $F_T$, image $F_I$, and discrepancy features $F_D$, we compute modality-specific weights using the following gating functions:

$$g_T = \sigma(W_T F_T), \quad g_I = \sigma(W_I F_I), \quad g_D = \sigma(W_D F_D). \tag{4}$$

where $W_T$, $W_I$, $W_D$ are trainable parameters and $\sigma$ denotes the sigmoid activation function. The final fused representation is:

$$F_{\text{fused}} = g_T \odot F_T + g_I \odot F_I + g_D \odot F_D. \tag{5}$$

To classify sarcasm, we utilize four independent classifiers for each modality-specific feature vector, including the fused representation. These logits are then concatenated to form a combined representation $logits_{\text{all}}$. Subsequently, the concatenated logits are passed through an MLP to produce the final prediction $P_{\text{final}}$.

### 2.5. Optimization Objective

GDCNet is trained to jointly optimize sarcasm classification and multimodal alignment. The objective consists of two components: a binary classification loss and a contrastive loss.

Sarcasm detection is formulated as a binary classification problem, and we compute the loss with cross-entropy between the predicted $P_{\text{final}}$ and the ground truth label $y$. Given a batch of $N$ training samples, the classification loss is computed as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log P_{\text{final},i} + (1-y_i)\log(1-P_{\text{final},i})\right]. \tag{6}$$

To enforce cross-modal consistency, we incorporate the contrastive loss $\mathcal{L}_{\text{cont}}$ (Eq. 2), which aligns paired image–text embeddings in the shared latent space.

The final objective combines both terms, with a hyperparameter $\alpha$ controlling the trade-off between classification accuracy and multimodal alignment:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \alpha\mathcal{L}_{\text{cont}}. \tag{7}$$

## 3. EXPERIMENTS

### 3.1. Experiments Details

***Dataset***: We evaluate our approach on MMSD2.0 [7], a refined and reliable benchmark for multimodal sarcasm detection. It is built upon MMSD [2] through the removal of spurious cues and the re-annotation of unreasonable samples, thereby offering a more robust basis for evaluation.
***Baselines***: To evaluate our approach, we compare it against competitive baselines in three categories. Text-modality methods include BiLSTM [14], SMSD [15], and BERT [16]; image-modality methods include ResNet [17] and ViT [18]; and multi-modality methods include InCrossMGs [19], HKE [5], Multi-view CLIP [7], DIP [20], TFCD [21], MOBA [22], CofiPara [10], and ADs [23].
***Implementation Details***: GDCNet is trained on four NVIDIA RTX 4090 GPUs using the CLIP backbone, with text and image feature dimensions of 512 and 768, respectively. For text generation, we employ LLaVA-Next-7B [9] as the MLLM. The model is optimized using Adam for 10 epochs with a batch size of 32, applying a learning rate of $5 \times 10^{-4}$ for most components and $1 \times 10^{-6}$ for CLIP. Weight decay of 0.05 and gradient clipping with a maximum norm of 5.0 are used for training stability. The contrastive loss employs a margin $m = 0.2$ and is weighted by $\alpha = 0.1$ in the final objective.

### 3.2. Main Results

As detailed in Table 1, our benchmark comparisons reveal critical insights into sarcasm detection and the efficacy of GDCNet.

Our GDCNet achieves state-of-the-art performance on MMSD2.0, consistently surpassing prior methods in both accuracy and F1-score. The superiority stems from GDCNet's novel modeling paradigm. Whereas CofiPara relies on joint text-image sarcasm rationale generation, our method isolates image description generation from textual

| **Method** | Acc.(%) | P(%) | R(%) | F1(%) |
|---|---|---|---|---|
| **Text-Only Methods** | | | | |
| BiLSTM [14] | 72.48 | 68.02 | 68.08 | 68.05 |
| SMSD [15] | 73.56 | 68.45 | 71.55 | 69.97 |
| BERT [16] | 76.52 | 74.48 | 73.09 | 73.78 |
| **Image-Only Methods** | | | | |
| ResNet [17] | 65.50 | 61.17 | 54.39 | 57.58 |
| ViT [18] | 72.02 | 65.26 | 74.83 | 69.72 |
| **Multi-Modal Methods** | | | | |
| InCrossMGs [19] | 79.83 | 75.82 | 78.01 | 76.90 |
| HKE [5] | 76.50 | 73.48 | 71.07 | 72.25 |
| Multi-view CLIP [7] | 85.14 | 80.33 | 88.24 | 84.09 |
| DIP [20] | 84.63 | 84.17 | 85.20 | 84.68 |
| TFCD [21] | 86.54 | 82.46 | 87.95 | 84.31 |
| MoBA [22] | 85.01 | 80.46 | 87.67 | 83.64 |
| CofiPara [10] | 85.66 | **85.79** | 85.43 | 85.61 |
| ADs [23] | 85.60 | 85.28 | 85.60 | 85.41 |
| **GDCNet (Ours)** | **87.38** | 83.39 | **89.51** | **86.34** |

**Table 1**. Performance comparison on MMSD2.0. Multimodal methods outperform unimodal ones, and GDCNet achieves the highest overall performance.

input, effectively mitigating textual bias. This isolation generates neutral visual observations, facilitating precise quantification of sarcasm through semantic and sentiment divergence analysis. Moreover, GDCNet introduces an adaptive gated fusion with explicit cross-modal divergence modeling, mitigating modality dominance and ensuring balanced multimodal integration.

### 3.3. Ablation Study

To assess the contribution of GDRM and its components, we conduct ablations on three configurations: removing the entire GDRM (*w/o GDRM*), and removing semantic (*w/o SemD*) or sentiment (*w/o SenD*) discrepancies. Results on MMSD2.0 are shown in Table 2.

Excluding GDRM leads to the greatest degradation, underscoring its necessity for explicit discrepancy modeling. Removing SemD weakens the model's ability to capture literal contradictions in incongruent image-text pairs, while removing SenD diminishes its capacity to detect subtle polarity shifts in text-driven sarcasm. Their complementarity enables the full model to maintain balanced performance and greater robustness.

| | Acc. (%) | P (%) | R (%) | F1 (%) |
|---|---|---|---|---|
| **Full Model** | **87.38** | **83.39** | **89.51** | **86.34** |
| -*w/o GDRM* | 84.42 | 78.56 | 86.17 | 82.19 |
| -*w/o SemD* | 86.23 | 80.27 | 87.09 | 83.54 |
| -*w/o SenD* | 85.98 | 81.74 | 87.63 | 84.58 |

**Table 2**. Ablation study on the MMSD2.0 dataset. Removing GDRM, SemD, or SenD results in clear performance drops, showing the importance of each component.

### 3.4. Comparison with LLM-based Methods

To further evaluate the performance of GDCNet, we conduct a comparative analysis against various direct LLM-based methods for sarcasm detection on the MMSD2.0 dataset. As presented in Table 3, we evaluate prominent MLLMs such as LLaVA[24], Qwen-VL[25], and GPT-4o[26], under both Zero-Shot and Chain-of-Thought (CoT) prompting strategies. As shown in Table 3, GDCNet consistently outperforms all LLM-based baselines in all metrics. While CoT improves reasoning for some models, MLLMs still struggle with sarcasm detection, underscoring the advantage of GDCNet's explicit discrepancy modeling for this challenging task.

| **Method** | Acc. (%) | P (%) | R (%) | F1 (%) |
|---|---|---|---|---|
| LLaVA (Zero-Shot) | 51.06 | 40.09 | 46.40 | 43.02 |
| LLaVA (CoT) | 48.69 | 40.93 | 65.17 | 50.28 |
| Qwen-VL (Zero-Shot) | 40.63 | 32.44 | 35.53 | 33.63 |
| Qwen-VL (CoT) | 58.86 | 56.82 | 58.67 | 57.26 |
| GPT-4o (Zero-Shot) | 71.07 | 79.52 | 71.07 | 70.24 |
| GPT-4o (CoT) | 74.26 | 65.81 | 72.68 | 68.92 |
| **GDCNet(Ours)** | **87.38** | **83.39** | **89.51** | **86.34** |

**Table 3**. Comparison results for LLM-based methods and GDCNet on the MMSD2.0 dataset.

### 3.5. MLLM Contribution Analysis

To investigate the impact of the MLLM on GDCNet's performance, we conducted an ablation study on the MMSD2.0 benchmark by substituting our image caption generator with BLIP-2[27] and LLaVA-NEXT[9]. For each MLLM, we generated image descriptions for all samples using an identical prompt, and then used them to train GDCNet with all other components fixed. As shown in Table 4, BLIP-2 is faster and more efficient, while LLaVA-NEXT produces captions with higher semantic consistency (CLIP-S), leading to better accuracy and F1. This highlights a trade-off between efficiency and caption quality, with the latter being more critical for sarcasm detection.

| **MLLM** | Time (s) | Tokens | CLIP-S | Acc. (%) | F1 (%) |
|---|---|---|---|---|---|
| BLIP-2 | **0.23** | 21.53 | 31.3 | 86.73 | 85.66 |
| LLaVA-NEXT | 1.70 | **67.29** | **49.2** | **87.38** | **86.34** |

**Table 4**. Performance and cost comparison of BLIP-2 and LLaVA-NEXT as caption generators in GDCNet.

## 4. CONCLUSION

In this paper, we propose GDCNet, a novel framework for multimodal sarcasm detection that effectively mitigates noise from LLM-generated sarcastic cues and explicitly models cross-modal incongruity. By leveraging LLM-generated image-grounded captions as cross-modal anchors, GDCNet performs modality-internal inconsistency detection, enabling finer-grained semantic and sentiment alignment. A gated fusion module that adaptively integrates visual, textual, and discrepancy signals is adopted, further alleviating modality dominance and reducing spurious correlations. Extensive experiments on benchmark datasets demonstrate that GDCNet consistently outperforms existing methods and achieves new state-of-the-art results on the MMSD2.0 benchmark. These results underscore the potential of leveraging LLMs not merely as data generators but as structural guides for capturing subtle cross-modal incongruities, advancing the frontier of complex multimodal understanding tasks such as sarcasm detection.

## 5. REFERENCES

[1] Raymond W Gibbs and Herbert L Colston, *Irony in language and thought: A cognitive science reader*, Psychology Press, 2007.

[2] Yitao Cai, Huiyu Cai, and Xiaojun Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 2506–2515.

[3] Xinyu Wang, Xiaowen Sun, Tan Yang, and Hongbo Wang, "Building a bridge: a method for image-text sarcasm detection without pretraining on image-text data," in *Proceedings of the first international workshop on natural language processing beyond text*, 2020, pp. 19–29.

[4] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu, "Multi-modal sarcasm detection via cross-modal graph convolutional network," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 1767–1777.

[5] Hui Liu, Wenya Wang, and Haoliang Li, "Towards multimodal sarcasm detection via hierarchical congruity modeling with knowledge enhancement," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 4995–5006.

[6] Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao, "Dynamic routing transformer network for multimodal sarcasm detection," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 2468–2480.

[7] Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu, "Mmsd2.0: Towards a reliable multimodal sarcasm detection system," *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10834–10845, 2023.

[8] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[9] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024.

[10] Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen, "Cofipara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9663–9687.

[11] Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri, "A survey of multimodal sarcasm detection," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024.

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26296–26306.

[13] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[14] J. Zhou, B. Xu, X. Xie, and Q. Xu, "Attention-based bidirectional lstm for text classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, pp. 1468–1477.

[15] Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang, "Sarcasm detection with self-matching networks and low-rank bilinear pooling," in *The World Wide Web Conference*, 2019, pp. 2115–2124.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

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

[19] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4707–4715.

[20] Changsong Wen, Guoli Jia, and Jufeng Yang, "Dip: Dual incongruity perceiving network for sarcasm detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2540–2550.

[21] Zhihong Zhu, Xianwei Zhuang, Yunyan Zhang, Derong Xu, Guimin Hu, Xian Wu, and Yefeng Zheng, "Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing," in *Proc. of IJCAI*, 2024.

[22] Yifeng Xie, Zhihong Zhu, Xin Chen, Zhanpeng Chen, and Zhiqi Huang, "Moba: Mixture of bi-directional adapter for multi-modal sarcasm detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, New York, NY, USA, 2024, MM '24, p. 4264–4272, Association for Computing Machinery.

[23] Soumyadeep Jana, Sahil Danayak, and Sanasam Ranbir Singh, "Ads: Adapter-state sharing framework for multimodal sarcasm detection," *arXiv preprint arXiv:2507.04508*, 2025.

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

[25] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023.

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al., "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.