GSDNet: Generative Sarcasm Discrepancy Network for Multimodal Sarcasm Detection

Anonymous ACL submission

Abstract

Multimodal sarcasm detection aims to identify sarcasm from given text-image pairs, where subtle contradictions between modalities are key to identifying irony. This task is essential for understanding nuanced human communications, especially in social media contexts. However, existing methods often overfit superficial textual patterns or fail to adequately model cross-modal incongruities, resulting in suboptimal performance. To address this, we propose the **G**enerative **S**arcasm **D**iscrepancy **N**etwork (GSDNet), which more effectively exploits cross-modal conflicts. GSDNet features a specialized **G**enerative **D**iscrepancy **R**epresentation Module (GDRM), which synthesizes image-aligned text using a large language model and quantifies both semantic and sentiment discrepancies by comparing the generated text with the original input. These discrepancies are then integrated with text and image representations via a gated fusion mechanism, enabling adaptive balancing of modality contributions and mitigating modality dominance and spurious correlations. Extensive experiments on two benchmarks demonstrate that GSDNet outperforms state-of-the-art models, achieving superior accuracy and robustness. These results highlight the effectiveness of discrepancy-based features and gated multimodal fusion in enhancing sarcasm detection.

1 Introduction

Sarcasm is a linguistic phenomenon in which the literal meaning of an utterance diverges significantly from its intended message. It is often employed to convey humor, criticism, or subtle social commentary and serves as a potent tool in human communication. Accurately detecting sarcasm is crucial for NLP tasks, such as sentiment analysis and opinion mining (Pang and Lee, 2008; Riloff et al., 2013), as it enables systems better to interpret the true sentiment behind seemingly contradictory expressions. In multimodal scenarios, sarcasm becomes even more complex. Given a text-image pair, the image may convey subtle visual cues that, when combined with the text, produce a sarcastic effect that goes beyond the literal meaning. **M**ultimodal **S**arcasm **D**etection (MSD) aims to classify whether a text-image pair indicates sarcasm.

With the rapid expansion of multimodal content on social media, MSD has emerged as a critical research area (Cai et al., 2019; Xu et al., 2020; Pan et al., 2020; Liang et al., 2021; Pramanick et al., 2022), which requires analyzing the interactions and contradictions between textual and visual cues. Recent approaches explore the relations among sarcasm cues from various perspectives, including attention mechanisms(Wang et al., 2020), graph-based modeling(Liang et al., 2022), external knowledge integration(Liu et al., 2022), and dynamic routing(Tian et al., 2023). These methods typically leveraged powerful language models like BERT (Devlin et al., 2019a) and RoBERTa (Liu et al., 2019) as backbones, constructing complex structured networks to model interactions across modalities. However, they often overfit domain-specific cues or rely heavily on superficial textual signals, limiting their ability to effectively capture the nuanced semantic and emotional incongruities characteristic of sarcasm.

The emergence of Large Language Models (LLMs) (Ouyang et al., 2022) and their multimodal LLMs (Zhu et al., 2023; Chen et al., 2023; Liu et al., 2023) have revolutionized natural language processing by providing unprecedented generative capabilities. Recent studies (Chen et al., 2024) have utilized these models to generate supplemented information to improve the generalization and interpretability of multimodal inputs. However, directly using LLMs to generate explanations for entire multimodal samples often leads to inconsistencies between the generated sarcastic interpretations and the actual sentiment or intent of the sample. These inconsistencies stem from the inherent complexity and variability of sarcasm, which involves subtle contradictions, contextual nuances, and implicit meanings that are challenging for LLMs to fully grasp. For instance, when prompted to explain why an image is sarcastic, LLMs may produce a wide range of interpretations due to the ambiguous nature of sarcasm. Under such circumstances, we hold that focusing solely on the image's core content leads to more consistent and accurate outputs by avoiding speculative reasoning about sarcasm.

Building on these insights, we propose a novel approach, **G**enerative **S**arcasm **D**iscrepancy **N**etwork (GSDNet), which leverages the generative capabilities of LLMs to facilitate the analysis of textual and visual cues. Specifically, a **G**enerative **D**iscrepancy **R**epresentation **M**odule (GDRM) is introduced to generate factual descriptions and contextual explanations by feeding only the image into the LLM. In this manner, we ensure that the generated text faithfully represents the visual content, while remaining unaffected by the original sarcastic text. We then quantify semantic and sentiment discrepancies by comparing the generated image descriptions with the original text, applying a text-image fidelity constraint to capture cross-modal incongruities. These discrepancies provide valuable features for sarcasm detection. Experimental results demonstrate that our approach avoids inconsistencies between LLMs-generated sarcastic interpretations and the actual intent of the sample by focusing on trustworthy image-aligned text. Our contributions are summarized as follows:

- We propose a novel **GSDNet**, the first framework for multimodal sarcasm detection that leverages trustworthy data generated by LLMs, effectively avoiding the biases and inconsistencies often introduced by complex samples and ensuring the reliability of the generated data.
- We introduce the **GDRM** to quantify semantic-sentiment gaps between generated visual descriptions and the original text, facilitating effective multimodal fusion for improved classification.
- Extensive experiments on two benchmark datasets show that GSDNet achieves **state-of-the-art** performance, significantly improving detection accuracy and generalization across various domains.

2 Related Work

2.1 Multimodal Sarcasm Detection

Sarcasm detection has traditionally focused on analyzing text to identify the contrast between literal and intended meanings (Tay et al., 2018; Babanejad et al., 2020). With the rise of social media, researchers began incorporating visual information to capture richer contextual cues. For example, Schifanella et al. (2016) first explored multimodal sarcasm detection by simply concatenating textual and visual embeddings, and Cai et al. (2019) later advanced the field by proposing a hierarchical fusion network and releasing the MMSD dataset.

Building on these early efforts, subsequent studies have improved multimodal sarcasm detection by better modeling the interplay between text and image. Approaches such as decomposition and relation networks (Xu et al., 2020), enhanced BERT-based methods with refined attention mechanisms (Pan et al., 2020; Wang et al., 2020), as well as graph neural networks (Liang et al., 2022) and optimal transport techniques (Pramanick et al., 2022) have been proposed to capture cross-modal features more effectively. Further enhancements include frameworks that incorporate external knowledge (Liu et al., 2022) and dynamic routing (Tian et al., 2023), while Qin et al. (2023) revealed that many existing models overly depend on superficial textual cues, prompting the development of refined benchmarks like MMSD2.0 and models based on vision-language pre-training such as CLIP (Radford et al., 2021). In summary, recent advances in multimodal sarcasm detection have focused on more effective integration of text and image features, addressing challenges like cross-modal alignment and external knowledge incorporation. Our approach further improves multimodal fusion by leveraging novel representation learning techniques to better capture the interplay between textual and visual cues, offering a more robust solution for sarcasm detection.

2.2 Multimodal Large Language Models

Multimodal large language models(MLLMs) have revolutionized natural language processing with unprecedented abilities in understanding and generating complex text (Brown et al., 2020; Ouyang et al., 2022). Their extension to multimodal data has further expanded application possibilities. Early works such as Frozen (Tsimpoukelli et al., 2021) and BLIP (Li et al., 2022) laid the groundwork by integrating visual encoders with language models, while subsequent approaches like BLIP2 (Li et al., 2023), MiniGPT4 (Zhu et al., 2023; Chen et al., 2023), LLaVA (Liu et al., 2023), and Qwen-VL (Bai et al., 2023) refined the alignment between visual and textual representations using adapter modules and efficient transformers.

Recently, an increasing number of studies have explored the use of MLLMs to enhance sarcasm detection. By leveraging the powerful generative capabilities of MLLMs in combination with visual inputs, these approaches offer more comprehensive representations that better capture the subtlety of sarcasm. CofiPara (Chen et al., 2024) leveraged MLLMs in a coarse-to-fine framework by generating rationales to guide sarcasm classification, thereby reducing noise and enhancing interpretability. Similarly, Jang and Frassinelli (2024) utilized MLLM-supported training on third-party labeled data to enhance model generalization in sarcasm detection. Tang et al. (2024) integrated visual instruction tuning with demonstration retrieval to construct instruction templates that boost out-of-domain performance.

Our approach exploits the generative and data augmentation capabilities of MLLMs to enrich multimodal representations and provide robust cues for sarcasm detection while mitigating the uncertainty often associated with LLM-generated outputs by leveraging trustworthy data.

3 Methodology

3.1 Problem Formulation

The task of multimodal sarcasm detection involves determining whether a given image-text pair (I, T) conveys sarcasm. This task can be formally defined as learning a classification function f that maps the image-text pair to a binary output $y \in \{0, 1\}$, where $y = 1$ indicates sarcasm. The primary challenge lies in capturing cross-modal incongruities, which refer to the subtle mismatches between the literal meaning of text and the contextual cues provided by the visual modality.

3.2 Model Framework

Traditional multimodal sarcasm detection methods often fuse image-text features directly, which risks overfitting to superficial textual patterns and overlooking nuanced cross-modal mismatches (Qin et al., 2023). To address this, we propose GSDNet, which introduces a generative discrepancy mechanism to enhance robustness. Instead of solely depending on direct feature fusion, GSDNet leverages LLMs to generate synthetic text $\hat{t}_i$, conditioned solely on the image v_i. By comparing this generated text with the original text t_i, the model can capture both semantic and emotional discrepancies, which serve as additional robust features for sarcasm classification.

The architecture of GSDNet is composed of three main components: 1) **Cross-modal Feature Alignment** , which involves extracting representations from both the image and text; 2) **Generative Discrepancy Representation Module**, which generates synthetic image description based solely on the image and computes the discrepancy between the generated text and the original text. 3) **Multimodal Fusion & Classification** fuses the extracted features, including those from the generative discrepancies, using gated networks, and classifies sarcasm based on these multimodal inputs.

By explicitly contrasting LLM-generated rationales with the original text, GSDNet reduces reliance on spurious textual correlations and strengthens contextual sarcasm reasoning. Subsequent sections detail each component.

3.3 Cross-modal Feature Alignment

Given a sample pair (I_i, T_i) from the dataset, where v_i represents the input image and t_i denotes the corresponding text, we utilize modality-specific encoders to extract high-dimensional embeddings. Specifically, the image encoder $E_v(\cdot)$ and the text encoder $E_t(\cdot)$ produce the following embeddings:

$$h_i^v = E_v(v_i), \quad h_i^t = E_t(t_i), \tag{1}$$

where $h_i^v \in \mathbb{R}^{d_v}$ and $h_i^t \in \mathbb{R}^{d_t}$ are the visual and textual feature representations, capturing the semantic and contextual information of the corresponding modalities.

To facilitate effective cross-modal alignment, we project the embeddings into a common latent space using learnable projection layers:

$$z_i^v = W_v h_i^v + b_v, \quad z_i^t = W_t h_i^t + b_t, \tag{2}$$

where W_v and W_t are the projection matrices, and b_v and b_t are the bias terms for the image and text modalities, respectively. The projected embeddings z_i^v and z_i^t share the same dimensionality d_z, promoting compatibility in the joint embedding space.

We adopt contrastive learning to align the projected features across modalities. Specifically, the

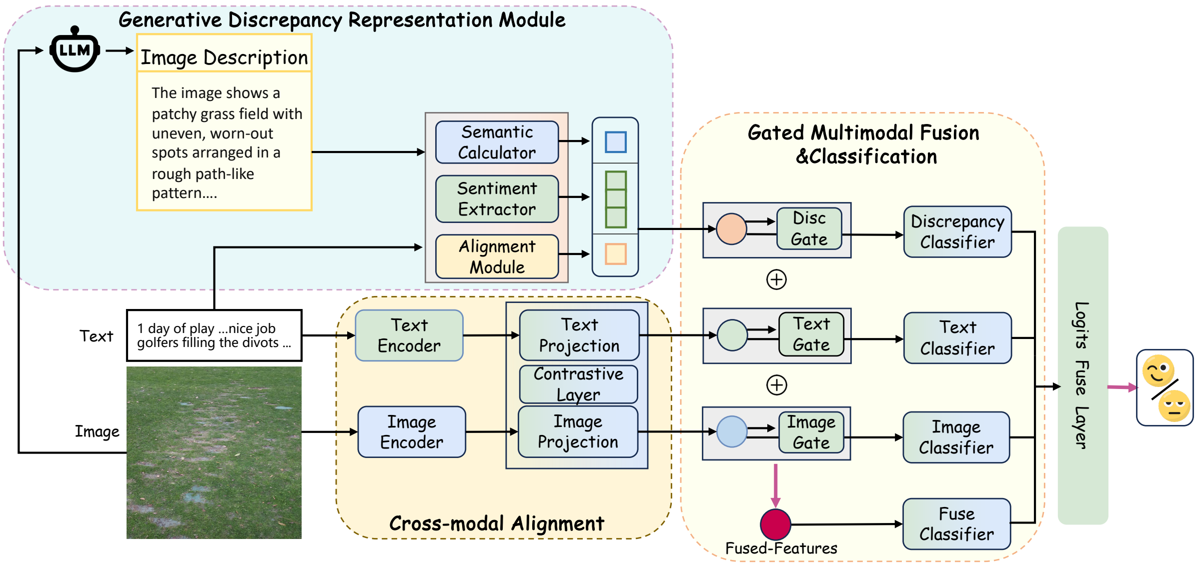


Figure 1: The Architecture of GSDNet. In the Gated Multimodel Fusion&Classifiction module, the four circles in different colors represent discrepancy features, text features, image features, and fused features, respectively.

cosine similarity between the paired embeddings is calculated as:

$$s_{ij} = \frac{z_i^v \cdot z_j^t}{\|z_i^v\| \|z_j^t\|}, \tag{3}$$

where s_{ij} denotes the similarity score between the i-th image and the j-th text. Positive pairs (i.e., matched image-text pairs) are encouraged to have higher similarity scores, while negative pairs (i.e., mismatched pairs) are pushed apart. The contrastive loss is defined as:

$$\mathcal{L}_{\text{cont}} = \sum_{i=1}^{B} \max\left(0,\ s_{ij \neq i} - s_{ii} + m\right). \tag{4}$$

This objective enforces alignment between semantically correlated visual and textual features, thereby enhancing cross-modal consistency and facilitating effective multimodal fusion in subsequent layers.

3.4 Generative Discrepancy Representation Module

The Generative Discrepancy Representation Module (GDRM) is designed to capture the implicit conflicts between the original text T and the image I by generating an unbiased textual description $\hat{T}$ using a LLM such as LLaVA-1.5(Liu et al., 2023). To maintain neutrality and avoid introducing sarcasm-related biases, the input to the LLM is strictly limited to the image content, excluding any multimodal sarcasm-labeled data. This ensures that $\hat{T}$ accurately reflects the visual semantics without being influenced by contextual sarcasm cues.

3.4.1 LLM-Based Text Generation

Given an input image I, the LLM generates a corresponding textual description $\hat{T}$ as follows:

$$\hat{T} = \text{LLM}(I), \tag{5}$$

where $\hat{T}$ is the generated text that aims to faithfully describe the visual content of the image. By avoiding the use of multimodal sarcasm-labeled data, this design ensures that $\hat{T}$ provides an unbiased and contextually neutral representation of the image content.

3.4.2 Discrepancy Computation

To quantify the inconsistency between the generated description $\hat{T}$ and the original text T, we compute three types of discrepancies: semantic discrepancy, emotional discrepancy, and visual-textual fidelity. These discrepancies collectively capture the underlying conflicts that may indicate sarcasm.

Semantic Discrepancy measures the divergence in meaning between the original text and the generated description. We use CLIP's text encoder to obtain the embeddings of both texts and calculate the cosine dissimilarity:

$$d_{sem} = 1 - \cos\left(z_T, z_{\hat{T}}\right), \tag{6}$$

where $z_T = \text{CLIP}_{\text{text}}(T)$ and $z_{\hat{T}} = \text{CLIP}_{\text{text}}(\hat{T})$ are the text embeddings of T and $\hat{T}$, respectively. A higher value of d_{sem} indicates a greater semantic divergence.

Sentiment Discrepancy captures shifts in sentiment between the original text and the generated description. Using a RoBERTa-based sentiment classifier(Ott et al., 2019), we obtain the sentiment probability distributions p_T and $p_{\hat{T}}$. The sentiment discrepancy is then calculated as:

$$d_{sen} = \left\| p_T - p_{\hat{T}} \right\|_1 , \tag{7}$$

where p_T and $p_{\hat{T}}$ are the sentiment distributions, representing the emotional intensities quantified by the sentiment analysis model. This metric effectively captures sentiment discrepancies, which are indicative of potential sarcasm.

Visual-Textual Fidelity evaluates the alignment between the LLM generated text and the corresponding image. We compute the cosine similarity between the image embedding and the generated text embedding:

$$d_v = \cos\left(z_I, z_{\hat{T}}\right) , \tag{8}$$

where $z_I = \text{CLIP}_{\text{image}}(I)$ is the image embedding. A lower value of d_v suggests that the generated text deviates from the visual content, indicating potential contextual conflicts.

3.4.3 Discrepancy Feature Representation

The computed discrepancies are concatenated to form the discrepancy feature vector:

$$D = [d_s, d_e, d_v]. \tag{9}$$

This vector is then projected through a multilayer perceptron (MLP) to obtain the final discrepancy representation:

$$F_{\text{D}} = \text{MLP}(D) \in \mathbb{R}^{d_f}, \tag{10}$$

where d_f is the dimensionality of the final feature vector. These discrepancy features are subsequently integrated into the sarcasm classification module, enriching the model's ability to detect nuanced incongruities.

3.5 Gated Multimodal Fusion & Classification

To optimally utilize textual, visual, and discrepancy-based features, we adapt the gated fusion mechanism. The mechanism 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:

$$\begin{aligned} g_T &= \sigma(W_T F_T), \\ g_I &= \sigma(W_I F_I), \\ g_D &= \sigma(W_D F_D), \end{aligned} \tag{11}$$

where W_T, W_I, and W_D are trainable parameters and σ denotes the sigmoid activation function. The final fused representation is then computed as:

$$F_{\text{fused}} = g_T \odot F_T + g_I \odot F_I + g_D \odot F_D. \tag{12}$$

$\odot$ denotes element-wise multiplication, which applies the weight to each corresponding element of the feature vector.

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 $\text{Logits}_{\text{all}}$. Subsequently, the concatenated logits are passed through an MLP to produce the final prediction:

$$P_{\text{final}} = \text{MLP}(\text{Logits}_{\text{all}}). \tag{13}$$

This hierarchical classification structure allows the model to effectively integrate information from all modalities while maintaining the interpretability of each individual feature vector's contribution.

3.6 Optimization Objective

The training process for GSDNet is designed to optimize sarcasm classification while ensuring effective multimodal representation learning. The overall objective consists of two main components: classification loss and contrastive loss. These loss functions work synergistically to direct the model's ability to effectively leverage both multimodal features and generative discrepancies, thereby enhancing the accuracy of sarcasm detection.

Sarcasm Classification Loss. Sarcasm detection is formulated as a binary classification problem, and we use **cross-entropy loss** to measure the difference between the predicted probability P_{final} and the ground truth label y. Given a batch of N training samples, the classification loss is computed as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log P_{\text{final},i}. \tag{14}$$

The loss encourages the model to learn optimal feature representations for sarcasm detection by assigning high confidence to correct predictions and penalizing incorrect classifications.

Contrastive Loss for Multimodal Alignment. To further enhance multimodal alignment, contrastive learning is employed to align the projected image and text embeddings in a shared latent space. The contrastive loss encourages this alignment with a margin $m = 0.2$. For each training batch, the contrastive loss $\mathcal{L}_{\text{cont}}$ is computed as shown in equation (4), which encourages the model to push the negative pairs further apart and pull the positive pairs closer in the shared latent space.

Final Objective Function. The final optimization objective combines the classification and contrastive losses, with the contrastive loss weighted by a hyperparameter α:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{cont}}. \tag{15}$$

Here, $\alpha = 0.1$ controls the trade-off between classification accuracy and multimodal alignment. By adjusting α, the model can prioritize one aspect over the other, ensuring that the final model incorporates both discrepancy-based features and multimodal fusion for optimal sarcasm detection.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate our approach on two widely-used datasets. MMSD (Cai et al., 2019) consists of image-text pairs collected from Twitter, which are randomly divided into training, validation, and test sets in the ratio of 80%, 10%, and 10%, respectively, serving as the primary benchmark for multimodal sarcasm detection. MMSD2.0 (Qin et al., 2023) is built upon MMSD and involves the removal of spurious cues and re-annotation of unreasonable samples, providing a more refined and reliable version for evaluation.

The details of two benchmarks, implementation details of the experiments, and evaluation metrics are reported in Appendix A.1, A.2, and A.3.

4.1.2 Baselines

To validate the effectiveness of our approach, we compare it with state-of-the-art baselines across three categories.

Text-modality methods include: 1) TextCNN (Kim, 2014), a convolutional neural network-based text classification model; 2) BiLSTM (Zhou et al., 2016), a bidirectional long short-term memory network for text classification; 3) SMSD (Xiong et al., 2019), a self-matching network utilizing low-rank bilinear pooling for sarcasm detection; 4) BERT (Devlin et al., 2019b) is a bidirectional transformer model pre-trained for language understanding.

Image-modality methods include: 1) ResNet (He et al., 2016), which utilizes image embeddings from the pooling layer for sarcasm classification; 2) ViT (Dosovitskiy et al., 2020), a vision transformer model pre-trained for visual representation learning.

Multi-modality methods include: 1) HFM (Cai et al., 2019), a hierarchical network that fuses multimodal features; 2) Att-BERT (Pan et al., 2020), which employs self-attention and co-attention mechanisms to capture intra- and inter-modality incongruity; 3) InCrossMGs (Liang et al., 2021), which captures sarcastic relations through in-modal and cross-modal graphs; 4) HKE (Liu et al., 2022), which incorporates external knowledge, such as image captions, using a hierarchical graph-based framework; 5) Multi-view CLIP (Qin et al., 2023) leverages CLIP's image-text interaction capabilities to fuse modality features; 6) LLaVA-1.5-7B (Liu et al., 2023) leverages a vision-language model with enhanced reasoning abilities for multimodal sarcasm detection; 7) DGLF (Zhu et al., 2024), a dual graph-based learning framework, uses a hypergraph for high-order relation modeling and a vanilla graph for high-frequency message propagation; 8) MOBA (Xie et al., 2024), a mixture of bi-directional adapters, dynamically integrates text and image features for sarcasm detection; 9) CofiPara-MSD (Chen et al., 2024) adopts a coarse-to-fine paradigm, leveraging LMMs for sarcasm reasoning and fine-tuning on target identification.

4.2 Main Results

As detailed in Table 1, our experiments reveal three critical insights about sarcasm detection through comprehensive benchmark comparisons.

The MMSD shows a significant modality discrepancy, as evidenced by comparative benchmarks: text-only method BERT achieves an accuracy of 83.60%, while image-only approaches such as ViT lag behind at 67.83%, confirming the inherent textual bias identified by Qin et al. (2023). The disparities across these four metrics highlight MMSD's inherent limitation in supporting cross-modal learning. In contrast, on the MMSD2.0 benchmark, the performance gap closes, with text-based and image-based methods achieving balanced results across all metrics.

Model	**MMSD**				**MMSD2.0**			
	Acc.(%)	P(%)	R(%)	F1(%)	Acc.(%)	P(%)	R(%)	F1(%)
Text-Only Methods								
TextCNN (Kim, 2014)	80.03	74.29	76.39	75.32	71.61	64.62	75.22	69.52
BiLSTM (Zhou et al., 2016)	81.90	76.66	78.42	77.53	72.48	68.02	68.08	68.05
SMSD (Xiong et al., 2019)	80.90	76.46	75.18	75.82	73.56	68.45	71.55	69.97
BERT (Devlin et al., 2019b)	83.60	78.50	82.51	80.45	76.52	74.48	73.09	73.78
Image-Only Methods								
ResNet (He et al., 2016)	64.76	54.41	70.80	61.53	65.50	61.17	54.39	57.58
ViT (Dosovitskiy et al., 2020)	67.83	57.93	70.07	63.40	72.02	65.26	74.83	69.72
Multi-Modal Methods								
HFM (Cai et al., 2019)	83.44	76.57	84.15	80.18	70.57	64.84	69.05	66.88
Att-BERT (Pan et al., 2020)	86.05	80.87	85.08	82.92	80.03	76.28	77.82	77.04
InCrossMGs (Liang et al., 2021)	86.10	81.38	84.36	82.84	79.83	75.82	78.01	76.90
HKE (Liu et al., 2022)	87.36	81.84	86.48	84.09	76.50	73.48	71.07	72.25
Multi-view CLIP (Qin et al., 2023)	88.22	82.03	88.13	84.97	85.14	80.33	88.24	84.09
LLaVA-1.5-7B (Liu et al., 2023)	-	-	-	-	85.18	**85.89**	85.20	85.11
DGLF (Zhu et al., 2024)	89.01	**84.96**	89.10	**86.98**	81.52	77.98	79.23	78.60
MoBA (Xie et al., 2024)	88.07	82.13	87.85	84.55	85.01	80.46	87.67	83.64
CofiPara-MSD (Chen et al., 2024)	88.46	83.46	88.26	85.79	85.66	85.79	85.43	85.61
GSDNet (**Ours**)	**89.17**	84.28	**89.67**	86.89	**87.38**	83.39	**89.51**	**86.34**

Table 1: Performance Comparison of Multimodal Sarcasm Detection Models. Baselines are divided into three categories, and the best value for each metric is highlighted in bold.

Our GSDNet achieves an accuracy of 87.38%, a recall of 89.51%, and an F1 of 86.34% on MMSD2.0, surpassing previous state-of-the-art methods including CofiPara-MSD and caption-enhanced LLaVA. Such superiority of GSDNet can stem from its fundamentally different modeling philosophy. Unlike CofiPara-MSD which feeds both text and image to LLMs for rationale generation, our method isolates image description generation to prevent textual bias propagation. This approach produces neutral visual observations that starkly contrast with sarcastic texts, enabling precise measurement of semantic contradictions and emotional mismatches. For instance, when processing an image of rusted pipes paired with text praising water quality, our model generates factual descriptions like "corroded plumbing components" to highlight incongruities, attaining higher precision than Multi-view CLIP in ambiguous cases.

The cross-modal analysis uncovers dataset limitations that hinder effective feature learning. Our method achieves state-of-the-art accuracy on both MMSD and MMSD2.0 datasets, demonstrating its capability to handle real-world scenarios with inherent modality imbalances. GSDNet addresses this through adaptive gated fusion and the robustness confirms that explicit cross-modal divergence modeling effectively mitigates modality dominance issues.

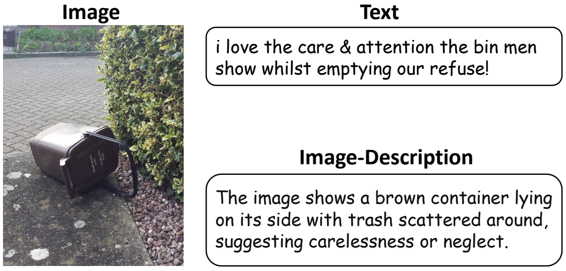
4.3 Ablation Study

We conduct systematic ablation experiments to evaluate the contribution of the Generative Discrepancy Representation Module (GDRM) through three configurations: the full model, removal of the entire GDRM module (w/o GDRM), and individual ablation of semantic/sentiment discrepancy pathways (w/o SemD and w/o SenD). Table 2 demonstrates that the complete GDRM architecture achieves optimal performance with 87.38% accuracy and 86.34% F1 score on MMSD2.0.

The removal of GDRM causes the most severe performance degradation, decreasing accuracy from 87.38% to 84.42% and F1 score from 86.34% to 82.19%. This substantial drop demonstrates the fundamental importance of cross-modal discrepancy modeling for sarcasm recognition. The precision suffers the largest reduction from 83.39% to 78.56%, indicating that GDRM effectively filters false positive predictions by detecting contradic-

Configuration	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 comparing different model variants on the MMSD2.0 dataset.



i love the care & attention the bin men show whilst emptying our refuse!

The image shows a brown container lying on its side with trash scattered around, suggesting carelessness or neglect.

Figure 2: A case identified as sarcastic by our model.

tions between modalities.

Partial ablations reveal asymmetric contributions of the two pathways. Disabling semantic discrepancy analysis significantly reduces F1 and recall, highlighting SemD's importance in detecting literal contradictions, such as incongruent image-text pairs with metaphorical expressions. In contrast, ablating the sentiment pathway primarily lowers precision, demonstrating SenD's effectiveness in capturing subtle emotional polarity shifts, particularly in text-dominated sarcasm.

The full model's balanced precision-recall profile outperforms all ablated variants, indicating synergistic interactions between the pathways. SemD establishes robust baseline detection by identifying explicit semantic contradictions, while SenD enhances performance by analyzing implicit emotional inconsistencies. This dual mechanism is especially effective for ambiguous cases where literal meaning aligns but emotional dissonance persists.

4.4 Case Study

To demonstrate GSDNet's effectiveness in detecting multimodal sarcasm, we analyze a representative example from the MMSD2.0 test set. The text exaggerates appreciation for garbage collectors: "*I love the care & attention the bin men show whilst emptying our refuse!*", The accompanying image in Figure 2 shows a tipped-over black container with scattered refuse, portraying disorder. While the text appears to praise the bin men's care, the image contradicts this, creating a sarcastic contrast between the positive language and the chaotic scene.

Our GSDNet uses GDRM to generate a description by processing only the image, rather than both image and text. It contrasts with methods that compare the modalities directly. As Figure 2 shows, GSDNet reduces the risk of unstable or inconsistent outputs that may arise from multimodal sarcasm cues affecting the LLM. This ensures more reliable image descriptions, forming a solid foundation for cross-modal discrepancy analysis.

The Semantic Discrepancy pathway captures the contrast between the text's praise for "care" and the disorder depicted in the image. A comparison between the generated description, "*The image shows a black container lying on its side with trash scattered around, suggesting carelessness or neglect.*" and the original text reveals a fundamental contradiction, thereby exposing the underlying irony. The Sentiment Discrepancy pathway further highlights the emotional contrast. The text conveys a positive sentiment, while the generated description implies a negative or indifferent tone. This emotional clash strengthens the sarcastic effect, reinforcing the discrepancy between exaggerated praise and a scene of neglect. This dual-pathway approach decodes the layered irony, demonstrating the necessity of generative discrepancy modeling for effective multimodal sarcasm detection.

5 Conclusion

Multimodal sarcasm detection requires nuanced modeling of cross-modal contradictions to decode implicit irony.ti In this paper, we present GSDNet, a framework that leverages generave discrepancy modeling to capture semantic and emotional conflicts between synthetic image descriptions and original text. Our experiments validate its superiority and effectiveness in handling complex multimodal sarcasm. By integrating gated fusion and contrastive alignment, the framework reduces reliance on biased textual cues and improves generalization. This work not only addresses key limitations in current MSD methods but also opens avenues for leveraging generative models to enhance multimodal understanding. While dependency on LLMs for text generation introduces potential instability, our approach provides a foundation for future research on interpretable and culturally adaptive sarcasm detection systems.

Limitations

Our approach inherits intrinsic constraints from its foundational architecture, with the framework's efficacy contingent upon LLM's ability to produce contextually aligned image descriptions. This dependency may compromise performance when analyzing abstract visuals (e.g., surreal artwork) or low-resolution imagery. Additionally, persistent challenges arise from cross-cultural divergences in sarcasm interpretation where discrepancies in symbolic visual metaphors intersect with the temporal evolution of linguistic irony, necessitating dynamic alignment between semantic parsing and pragmatic contexts. Collectively, these constraints underscore the imperative for multimodal evaluation protocols that integrate human expertise to strengthen contextual adaptability across linguistic and cultural boundaries.

References

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. 2024. 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)*, pages 9663–9687, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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

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*, pages 770–778.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

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, Doha, Qatar. Association for Computational Linguistics.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900. PMLR.

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*, pages 4707–4715.

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

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

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

Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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 the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940.

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

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*, pages 8748–8763.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 704–714.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 1136–1145.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1732–1742, Mexico City, Mexico. Association for Computational Linguistics.

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

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. 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)*, pages 2468–2480, Toronto, Canada. Association for Computational Linguistics.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Xinyu Wang, Xiaowen Sun, Tan Yang, and Hongbo Wang. 2020. 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*, pages 19–29, Online. Association for Computational Linguistics.

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

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

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*, pages 3777–3786, Online. Association for Computational Linguistics.

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

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

Zhihong Zhu, Kefan Shen, Zhaorun Chen, Yunyan Zhang, Yuyan Chen, Xiaoqi Jiao, Zhongwei Wan, Shaorong Xie, Wei Liu, Xian Wu, and Yefeng Zheng. 2024. DGLF: A dual graph-based learning framework for multi-modal sarcasm detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2900–2912, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Dataset Details

Datasets	Split	Positive	Negative	Total	Avg Len
MMSD	Training	8,642	11,174	19,816	15.71
	Validation	959	1,451	2,410	15.72
	Test	959	1,450	2,409	15.89
MMSD2.0	Training	9,576	10,240	19,816	13.42
	Validation	1,042	1,368	2,410	13.64
	Test	1,037	1,372	2,409	13.52

Table 3: Dataset structure and details

As shown in Table 3, two datasets, MMSD and MMSD2.0, are presented. For the MMSD dataset, the training set contains 19,816 samples, with 8,642 positive and 11,174 negative ones. The validation and test sets each have 2,410 and 2,409 samples respectively. The average lengths of samples in these splits are around 15.7 - 15.9. For MMSD2.0, the training set also has 19,816 samples (9,576 positive and 10,240 negative). The validation and test sets have 2,410 and 2,409 samples respectively, with average sample lengths of approximately 13.4 - 13.6.

A.2 Implementation Details

Our model is trained on four RTX 4090 GPUs using the CLIP backbone, with text and image feature dimensions setting to 512 and 768, respectively. We employ LLaVA-Next-8B(Liu et al., 2024) as the MLLM for text generation. Training is conducted over 10 epochs with a batch size of 32, using the Adam optimizer with learning rates of 5e-4 for all modules except CLIP, which uses 1e-6. To enhance generalization, weight decay is set to 0.05, and gradient clipping is applied with a max grad norm of 5.0, ensuring stable and efficient training.

A.3 Evaluation Metrics

We evaluate the model's performance using four key metrics: Accuracy (**Acc.**), Precision **(P)**, Recall **(R)**, and F1 Score (**F1**). Accuracy measures the overall correctness of predictions. Precision quantifies the proportion of correctly predicted sarcastic instances among all instances predicted as sarcastic, while Recall measures the proportion of correctly predicted sarcastic instances among all actual sarcastic instances. The F1 Score, derived as the harmonic mean of Precision and Recall, provides a balanced measure of the model's performance.

A.4 LLM Prompt Template

The prompt is as follows:

“Please generate a text description of the image I provide, focusing on the main content within 77 words. Include details about any people, main subjects, environment, and any hidden emotions or feelings that the image might convey. Please notice that the generation should be less than 77 tokens. Image:{T}.”
{T} is the corresponding image.