Leveraging Generative Large Language Models with Visual Instruction and Demonstration Retrieval for Multimodal Sarcasm Detection

Binghao Tang, Boda Lin, Haolong Yan, Si Li*
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{tangbinghao,linboda,yanhaolong,lisi}@bupt.edu.cn

Abstract

Multimodal sarcasm detection aims to identify sarcasm in the given image-text pairs and has wide applications in the multimodal domains. Previous works primarily design complex network structures to fuse the image-text modality features for classification. However, such complicated structures may risk overfitting on in-domain data, reducing the performance in out-of-distribution (OOD) scenarios. Additionally, existing methods typically do not fully utilize cross-modal features, limiting their performance on in-domain datasets. Therefore, to build a more reliable multimodal sarcasm detection model, we propose a generative multimodal sarcasm model consisting of a designed instruction template and a demonstration retrieval module based on the large language model. Moreover, to assess the generalization of current methods, we introduce an OOD test set, RedEval. Experimental results demonstrate that our method is effective and achieves state-of-the-art (SOTA) performance on the in-domain MMSD2.0 and OOD RedEval datasets. The source code and RedEval are available at https://github.com/TangBinghao/naacl2024.

1 Introduction

Sarcasm is a linguistic phenomenon of verbal irony where the literal meaning contradicts the real intent of the speaker. Sarcasm detection aims to identify the actual sentiment of the user and can be widely applied in various scenarios such as public opinion mining (Pang et al., 2008; Riloff et al., 2013) and social media analysis (Tsur et al., 2010). Recently, due to the rapid surge of multimodal data on social media, multimodal sarcasm detection has gained increasing attraction and significance (Cai et al., 2019; Xu et al., 2020; Pan et al., 2020; Wang et al., 2020; Liang et al., 2021, 2022; Pramanick et al., 2022; Liu et al., 2022a; Tian et al., 2023; Qin et al., 2023). As shown in Figure 1, the given image-text pair is sarcastic because the image fails to show a good execution of "replacing grass" while the text describes it as a "wonderful job".

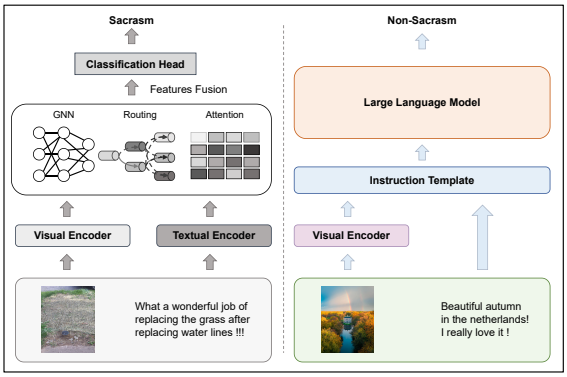

Figure 1: Illustration of multimodal sarcasm detection. The left of the figure conveys a sarcastic meaning with the contrast between "grass replacing" and "wonderful job". The right part displays an image of "beautiful autumn", which is semantically consistent and shows no sarcasm. Previous works rely on complex model structures for feature fusion followed by classification, whereas our method generates answers based on LLMs.

Previous studies on multimodal sarcasm detection capture the sarcasm cues of multimodal contents from different perspectives, such as attention-based methods (Wang et al., 2020; Pan et al., 2020), graph-based methods (Liang et al., 2021, 2022), extra knowledge enhancement (Liu et al., 2022a), and dynamic routing (Tian et al., 2023). Those methods primarily rely on BERT (Kenton and Toutanova, 2019) or RoBERTa (Liu et al., 2019) models, constructing complex structured networks to model features across two modalities. Despite their effectiveness, a notable concern arises about the tendency to overfit specific in-domain data features, which may hinder the generalization of models. Furthermore, Qin et al. (2023) points out that existing models may rely too heavily on spurious textual cues, which can decrease the utilization of

*Corresponding author

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
June 16-21, 2024

cross-modal features and limit the performance in in-domain situations. However, effectively balancing and integrating cross-modal feature interactions remains a critical challenge in enhancing the generalization and robustness of multimodal sarcasm detection models.

Fortunately, significant progress has been made in various NLP generation tasks with the development of Large Language Models (LLMs) (Ouyang et al., 2022). Further leveraging LLMs and extending them to multimodal domains, the Multimodal Large Language Models (MLLMs) (Zhu et al., 2023; Chen et al., 2023; Liu et al., 2023a,b) have also significantly improved various multimodal tasks and show great generalization.

Therefore, to build a more reliable multimodal sarcasm detection model, we redefine multimodal sarcasm detection as a generative task to take advantage of the powerful MLLMs. To further leverage and enhance the performance of MLLMs, we design a detailed instruction template and propose a simple yet effective demonstration retrieval module. Furthermore, considering the lack of research assessing the generalization of current multimodal sarcasm detection models, we collect multimodal data from other social media platforms and propose a new test set named RedEval for OOD evaluation. In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to attempt to explore the generalization of multimodal sarcasm models and the first to redefine multimodal sarcasm detection as a generation task utilizing MLLMs for better balance and integration of cross-modal interactions.
- We propose a new test dataset, RedEval, comprising image-text pairs collected from other social media, to assess the generalization of existing models, aiming to construct more reliable multimodal sarcasm models.
- We design an instruction template and a retrieval module to further enhance our generative multimodal sarcasm model. Experimental results on MMSD2.0 and RedEval demonstrate that our method is effective in both in-domain and OOD situations and achieves SOTA performance.

2 Related Work

2.1 Multimodal Sarcasm Detection

Traditional sarcasm detection task aims to identify the sentiments of users and detect the presence of sarcasm from textual modality (Zhang et al., 2016; Tay et al., 2018; Babanejad et al., 2020). Due to the surge of multimodal data in social media, multimodal sarcasm detection has gradually attracted much attention.

Schifanella et al. (2016) firstly explores the multimodal sarcasm detection task by concatenating the textual and visual embeddings. Cai et al. (2019) proposes a hierarchical fusion network and releases a multimodal public dataset, i.e., MMSD. Subsequent studies further model the commonalities and incongruity between visual and textual modalities by a decomposition and relation network (Xu et al., 2020), BERT-based (Kenton and Toutanova, 2019) models through modified attention mechanisms (Pan et al., 2020; Wang et al., 2020), graph neural networks (Liang et al., 2021, 2022) and optimal transport (Pramanick et al., 2022). And Liu et al. (2022a) proposes a hierarchical framework with external knowledge enhancement for multimodal sarcasm detection. Recently, Tian et al. (2023) applies a dynamic routing network to model the cross-modal incongruity. Furthermore, Qin et al. (2023) discovers that existing models may overly rely on spurious textual cues rather than cross-modal features. This leads to the introduction of a new benchmark, MMSD2.0, and the proposal of a novel framework based on the vision-language pre-trained model CLIP (Radford et al., 2021) to capture sarcasm cues from diverse perspectives. Compared with prior works, our method redefines multimodal sarcasm detection as a generative task.

2.2 Multimodal Large Language Models

Large Language Models (LLMs) have achieved widespread success in the field of NLP. From early-stage models like BERT (Kenton and Toutanova, 2019) and GPT-2 (Radford et al., 2019) to more recent GPT-3 (Brown et al., 2020), instructGPT (Ouyang et al., 2022), and various other open-source large-scale language models, such as LLaMA (Touvron et al., 2023a) and LLaMA2 (Touvron et al., 2023b), there has been substantial development in the field of NLP, particularly in the area of natural language understanding and generation.

In the research of multi-modality, how to apply those powerful LLMs to multimodal tasks

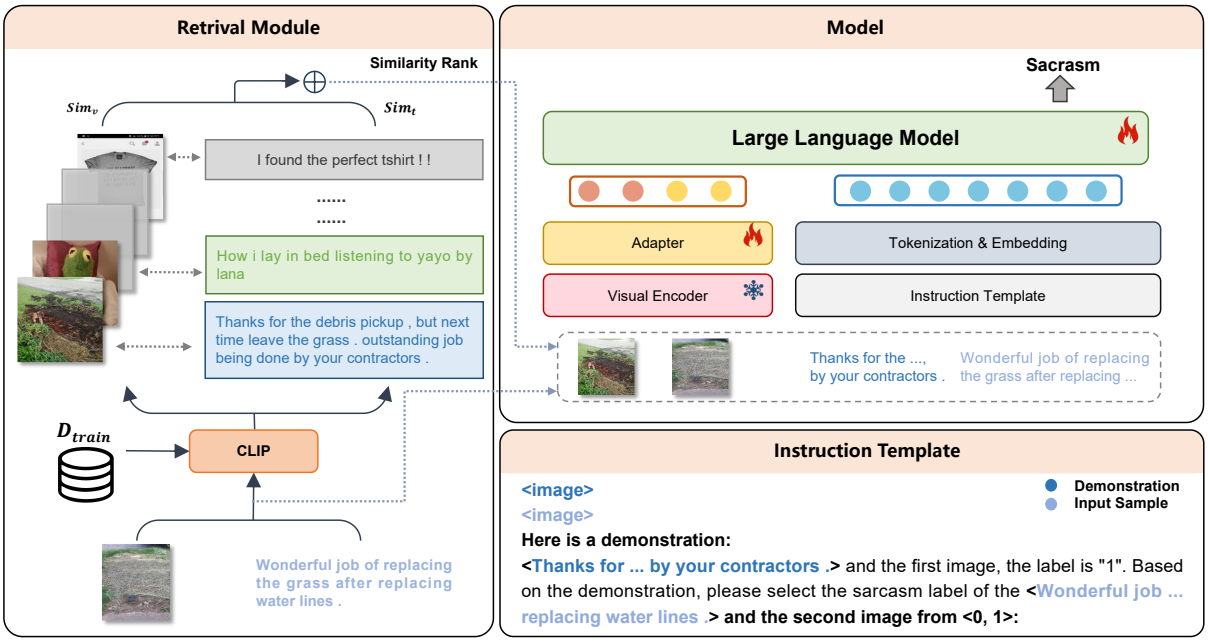

Figure 2: The overview of our model, which includes the retrieval module, instruction template, and generative multimodal large language model. The demonstration and sample images and texts are processed through a visual encoder with an adapter and the instruction template before being input together into the large language model.

has also gradually gained significant attraction. Early research like Frozen (Tsimpoukelli et al., 2021), achieves impressive performance by training a visual encoder to encode the image input as a prefix in a frozen pre-trained language model. BLIP (Li et al., 2022) pre-trains a multimodal mixture of encoder-decoder model to further boost vision-language tasks. BLIP2 (Li et al., 2023) proposes a Q-former to efficiently align visual features to LLMs. Additionally, other studies such as MiniGPT4 (Zhu et al., 2023; Chen et al., 2023), LLaVA (Liu et al., 2023a,b) and Qwen-VL (Bai et al., 2023) employ an adapter like a linear layer or multi-layer perceptron to further align the image features extracted from visual encoders like ViT (Dosovitskiy et al., 2020). We apply the multimodal sarcasm detection task to the generative framework of multimodal large language models to address the problems of insufficient generalization and inadequate reliance on multimodal features.

3 Methodology

In this section, we present the overview of our method. We first present the brief task formulation and describe the MLLM-based generative framework. Then we detail our retrieval module and introduce the training and generation process.

3.1 Task Formulation

Given image-text pairs $\langle v_i, t_i \rangle$, where v_i is the i-th image input and t_i is the i-th text input. MLLM needs to generate the sarcasm label from the label set $S = \{\epsilon_p, \epsilon_n\}$ based on v_i and t_i, where ϵ_p and ϵ_n are the positive and negative labels.

3.2 Model Framework

For the generative MLLMs, we leverage LLaVA-1.5 (Liu et al., 2023a) as our backbone. LLaVA-1.5 adopts a multi-layer perceptron as the cross-modal projection to connect the vision encoder and large language model. Liu et al. (2023a) further pre-trains the vision-language connector on the 600K public image-text pairs instructions data, which shows the strong power of various multimodal tasks. As shown in Figure 2, given the image-text pair, we first retrieve the best demonstration from the training set. Then we obtain the visual features of both the demonstration and sample images using the visual encoder and adapter. For the demonstration text and sample text, we input them into LLM along with the visual features in the format of an instruction template.

3.3 Retrieval Module

To better prompt MLLMs to generate the right answers, we introduce a retrieval module for MLLMs to search for demonstrations, aiming at further

bridging the gap between MLLMs and the specific multimodal sarcasm detection task.

As shown in Figure 2, for the given image-text pairs $\langle v_i, t_i \rangle$, we first obtain their corresponding embeddings by CLIP (Radford et al., 2021):

$$\text{Emb}_v(i) = \text{CLIP}_{\text{vis}}(v_i) \quad (1)$$

$$\text{Emb}_t(i) = \text{CLIP}_{\text{text}}(t_i) \quad (2)$$

where CLIP_{vis} and $\text{CLIP}_{\text{text}}$ are the visual and textual encoder of CLIP.

For each sample $\langle \text{Emb}_v(i), \text{Emb}_t(i) \rangle$, we calculate the cosine similarity of image and text modalities separately with the samples in the training set D_{train}:

$$Sim_v(i) = \frac{\text{Emb}_v(i) \cdot \mathbb{V}}{|\text{Emb}_v(i)||\mathbb{V}|} \quad (3)$$

$$Sim_t(i) = \frac{\text{Emb}_t(i) \cdot \mathbb{T}}{|\text{Emb}_t(i)||\mathbb{T}|} \quad (4)$$

where $\mathbb{V}$ and $\mathbb{T}$ are the image and text embeddings from the training set D_{train}.

Finally, we select the sample with the highest average similarity score as the corresponding demonstration:

$$Demon(i) = \arg\max \frac{Sim_v(i) + Sim_t(i)}{2} \quad (5)$$

where $Demon(i)$ is the similarity score of demonstration of i-th image-text sample. Then the retrieved demonstration and the sample, after being processed through the instruction template, are input into LLM together, as shown in Figure 2.

3.4 Optimization Objective

Consistent with the loss calculation in auto-regressive LLMs, we only compute the cross-entropy loss for the response of MLLM, i.e., the label of the image-text pair:

$$\mathcal{L} = \sum_{i=1}^{n} -\log p_\theta(\epsilon_i | \text{instruction}_i) \quad (6)$$

where instruction_i is the i-th instruction containing the information of image-text pair, ϵ_i is the corresponding predicted label and θ represents the parameters of the MLLM.

3.5 Constrained Decoding

For the generative MLLMs, the results may not fully comply with the requirements even when the output format is specified in the input instruction. This poses a challenge for the classification results of the multimodal sarcasm detection task. To address this issue, we implement constrained decoding (De Cao et al., 2020), ensuring that the model can only generate outputs based on the label set.

4 Experiments

4.1 Dataset

In-Domain Dataset We evaluate our method on MMSD (Cai et al., 2019) and MMSD2.0 (Qin et al., 2023) datasets. The MMSD dataset originates from the image-text pairs collected by Cai et al. (2019) on Twitter[1] and is randomly divided into training, validation, and test sets in the ratio of 80%, 10%, and 10% respectively. The MMSD2.0 (Qin et al., 2023) dataset is built upon MMSD, involving the removal of spurious cues and re-annotating unreasonable samples on the textual content. The statistics of the MMSD and MMSD2.0 datasets are shown in Table 1. For a fair comparison, we conduct the same data preprocessing on the MMSD dataset following previous works.

Out-of-Domain Dataset To assess the generalization of current multimodal sarcasm models, we propose a new test dataset called RedEval. Considering that the existing image-text pairs in the MMSD and MMSD2.0 datasets are all from the same social media Twitter, we select image-text pairs from another social media platform Reddit[2] as the out-of-distribution data. Specifically, we select image-text data from the "sarcasm" subreddit as positive sarcasm samples, and a certain number of samples from other subreddits such as "aww", "funny", "pics", and "popular" as the non-sarcastic samples. Following Qin et al. (2023), we remove the emotions from the data. We also employ 3 graduate students to ensure the quality of the image-text pairs in RedEval aligns with the intended meaning of sarcastic labels. In detail, each image-text pair is reviewed by two annotators. They are required to predict the sarcasm label. Only samples that receive consistent predictions from two annotators are left. Moreover, we let another annotator predict the sarcasm labels and then calculated the Kappa coefficient with the gold labels. We obtained a Kappa value of 0.793, indicating a high level of consistency. The statistics of RedEval are shown in Table 2, the maximum and average lengths are different from MMSD and MMSD2.0.

[1] https://twitter.com/

[2] https://www.reddit.com/

Datasets	Split	Positive	Negative	Total	Max Len	Min Len	Avg Len
MMSD	Training	8,642	11,174	19,816	70	1	15.71
	Validation	959	1,451	2,410	55	1	15.72
	Test	959	1,450	2,409	64	1	15.89
MMSD2.0	Training	9,576	10,240	19,816	66	1	13.42
	Validation	1,042	1,368	2,410	55	4	13.64
	Test	1,037	1,372	2,409	52	4	13.52

Table 1: The statistics of MMSD and MMSD2.0. Len denotes the number of words in the corresponding dataset.

	Pos.	Neg.	Total	Max	Min	Avg
RedEval	395	609	1,004	54	1	7.35

Table 2: The statistics of RedEval. Pos. and Neg. are the positive and negative samples. Max, Min, and Avg are the number of words as mentioned in Table 1.

4.2 Experimental Settings

Based on LLaVA-1.5-7B (Liu et al., 2023a), we use " CLIP-ViT-L-336px" as the vision encoder and "Vicuna-v1.5-7B" as the LLM. We use the same vision encoder to obtain image and text embeddings in the retrieval module. We utilize "BLIP2-FlanT5-XL" to obtain image captions. Given the limitations of task-specific data and computational resources, we choose Parameter-Efficient-Fine-Tuning (PEFT) for the training stage. Specifically, we adopt LoRA (Hu et al., 2021) and inject the low-rank matrices as adapters into MLLM. The rank of the update matrices is 128 and the scaling factor of LoRA is 256. We freeze the vision encoder and fine-tune the vision-language connector and LLM following Liu et al. (2023a). The learning rate for the vision-language connector is 2e-5 and the learning rate for LLM is 2e-4. The batch size is 12 and the training epoch is 5. We adopt the constrained greedy search for inference. All models are trained on 2 NVIDIA 3090Ti GPUs and tested on a single NVIDIA 3090Ti GPU.

4.3 Baselines

Following prior works, we compare our method with unimodal and multimodal baselines for multimodal sarcasm detection on MMSD and MMSD2.

Text-Modality Methods (1) TextCNN (Kim, 2014) is a text classification network based on the convolutional neural network. (2) BiLSTM (Zhou et al., 2016) is a bi-directional long short-term memory network for text classification. (3) SMSD (Xiong et al., 2019) is a self-matching network with low-rank bilinear pooling for sarcasm detection. (4) RoBERTa (Liu et al., 2019) is a robustly optimized BERT (Kenton and Toutanova, 2019) pre-trained language model. (5) ChatGLM2-6B (Du et al., 2022) is an open bilingual language model based on the general language model framework, with 6.2 billion parameters. (6) LLaMA2-7B (Touvron et al., 2023b) is a foundation LLM pre-trained on 2 trillion tokens, with 7 billion parameters. We refer to ChatGLM2-6B and LLaMA2-7B as the LLM-based methods.

Image-Modality Methods. (1) ResNet (He et al., 2016) utilizes the image embedding that is produced by the pooling layer to detect sarcasm. (2) ViT (Dosovitskiy et al., 2020) is a pre-trained vision transformer model.

Multi-Modality Methods. (1) HFM (Cai et al., 2019) is a hierarchical network with multimodal fusion. (2) D&R Net (Xu et al., 2020) proposes a decomposition and relation network to model the relationship between image and text. (3) Att-BERT (Pan et al., 2020) adopts self-attention and co-attention mechanisms to model the intra-modality and inter-modality incongruity respectively. (4) InCrossMGs (Liang et al., 2021) utilizes in-modal and cross-modal graphs to capture sarcastic relations between two modalities. (5) CMGCN (Liang et al., 2022) proposes a fine-grained cross-modal graph architecture to capture sarcastic clues. (6) HKE (Liu et al., 2022a) uses a hierarchical graph-based framework and incorporates external knowledge like image captions for multimodal sarcasm detection. (7) DynRT-Net (Tian et al., 2023) proposes a dynamic routing transformer network to capture the sarcastic clues from images and texts. (8) Multi-view CLIP (Qin et al., 2023) utilizes a framework based on CLIP (Radford et al., 2021) from image view, text view, and image-text interactions view for multimodal sarcasm detection, which is current State-Of-The-Art (SOTA) multimodal sarcasm model. (9) LLaVA1.5-7B (Liu et al., 2023a) adopts a multi-

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
RoBERTa (Liu et al., 2019)*	93.97	90.39	**94.59**	92.45	79.66	76.74	75.70	76.21
ChatGLM2-6B (Du et al., 2022)	94.02	93.46	94.14	93.76	78.41	78.15	78.65	78.23
ChatGLM2-6B (Du et al., 2022)$^{\nu}$	94.02	93.46	94.14	93.76	80.08	80.52	81.04	80.04
LLaMA2-7B (Touvron et al., 2023b)	93.97	93.42	94.09	93.72	82.52	82.15	82.46	82.27
LLaMA2-7B (Touvron et al., 2023b)$^{\nu}$	**94.02**	**93.46**	94.14	**93.76**	**84.68**	**84.40**	**84.94**	**84.53**
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
D&R Net (Xu et al., 2020)*	84.02	77.97	83.42	80.60	–	–	–	–
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	–	–	–	–
CMGCN (Liang et al., 2022)*	86.54	–	–	82.73	79.83	75.82	78.01	76.90
HKE (Liu et al., 2022a)*	87.36	81.84	86.48	84.09	76.50	73.48	71.07	72.25
DynRT-Net (Tian et al., 2023)	93.59	93.06	**93.60**	93.31	71.40	71.80	72.17	71.34
Multi-view CLIP (Qin et al., 2023)*	88.33	82.66	88.65	85.55	85.64	80.33	**88.24**	84.10
LLaVA1.5-7B (Liu et al., 2023a)	**93.67**	**93.70**	93.14	**93.40**	85.18	85.89	85.20	85.11
Ours	89.97†	89.26†	89.58†	89.42†	**86.43**†	**87.00**†	86.30	**86.34**†

Table 3: Experimental results on MMSD and MMSD2.0. The best results are highlighted in bold, and the second-best results are underlined. $*$ denotes the experimental results from Qin et al. (2023). ν denotes that the text-modality method takes the image captions as visual information inputs. † means our method outperforms Multi-view CLIP significantly with $p < 0.001$. Compared with MMSD2.0, the performance of text-modality methods reaches the highest on MMSD, indicating that MMSD is not sufficient to measure the effectiveness of multimodal methods.

layer perceptron as an adapter to connect the vision encoder and LLM, which has 7 billion parameters. It is given the image-text pairs and required to predict the labels. LLaVA1.5-7B is our base model.

For the out-of-domain situation, we compare our method with ChatGLM2-6B, LLaMA2-7B, DynRT-Net, Multi-view CLIP, and LLaVA1.5-7B. These models are all trained on the training set of MMSD and MMSD2.0 and tested on RedEval.

4.4 Main Results

Following Qin et al. (2023), we adopt accuracy (Acc.), macro-average precision (P), macro-average recall (R), and macro-average F1 score (F1) as metrics to assess the performance of our model.

Datasets Discussion. As shown in Table 3, for the MMSD dataset, the performance of LLMs like ChatGLM2-6B and LLaMA2-7B in the text modality methods reaches a relatively high level, even outperforming the multimodal methods. It is consistent with the experimental result of RoBERTa reported in Qin et al. (2023). This suggests that there indeed exists a problem with the text modality data in MMSD, which undermines the dependency of multimodal methods on image-text modality features. Furthermore, we observe that the performance of ChatGLM2-6B, LLaMA2-7B, and LLaVA1.5-7B on MMSD is very similar. This suggests that the performance of models on MMSD may have already reached the upper limit, making further improvements challenging. As for the MMSD2.0 dataset, multimodal model approaches generally outperform unimodal methods. Also, LLM-based methods utilizing image captions as visual information inputs could achieve better performance. This indicates that MMSD2.0 strengthens the dependency on cross-modal features, preventing models from relying solely on textual infor-

Model	MMSD for OOD				MMSD2.0 for OOD			
	Acc. (%)	**P (%)**	**R (%)**	**F1 (%)**	**Acc. (%)**	**P (%)**	**R (%)**	**F1 (%)**
Text-Only Methods								
ChatGLM2-6B (Du et al., 2022)	58.47	46.51	49.13	41.12	77.19	76.74	74.57	75.22
ChatGLM2-6B (Du et al., 2022)$^{\nu}$	58.57	46.79	49.21	41.17	79.28	78.92	80.25	78.95
LLaMA2-7B (Touvron et al., 2023b)	58.57	46.79	49.21	41.17	79.48	78.66	79.66	78.93
LLaMA2-7B (Touvron et al., 2023b)$^{\nu}$	58.67	47.35	49.34	41.40	81.38	80.47	80.60	80.53
Multi-Modal Methods								
DynRT-Net (Tian et al., 2023)	58.57	47.06	49.25	41.35	74.80	75.58	76.69	74.66
Multi-view CLIP (Qin et al., 2023)	**76.29**	**75.67**	**73.70**	**74.30**	80.98	80.85	**82.62**	80.73
LLaVA-1.5-7B	<u>61.25</u>	<u>52.63</u>	<u>57.68</u>	<u>47.13</u>	<u>82.77</u>	**83.66**	82.25	<u>82.44</u>
Ours	59.16	49.70	48.67	41.47	**83.47**†	<u>83.12</u>	<u>82.60</u>	**82.83**†

Table 4: Experimental results on RedEval. The best results are highlighted in bold, and the second-best results are underlined. The models are trained on the training set of MMSD and MMSD2.0 and tested on RedEval. ν and † denotes as the same in Table 3.

mation to predict the correct labels. In summary, MMSD falls short in evaluating current multimodal methods, whereas MMSD2.0 offers a more effective assessment.

Additionally, we observe that the performance of LLMs on MMSD2.0 can reach a relatively high standard compared to current multimodal methods. Yet the choice of base models is important and influences the final performance. For example, the base performance of LLaMA2-7B is higher than that of ChatGLM2-6B. We also observe that MLLM like LLaVA outperforms LLM-based methods. This indicates that connecting a visual encoder and LLM through an adapter could process a more diverse range of image information, leading to better performance in detecting multimodal sarcasm. In contrast, using LLMs alone or converting images into captions for inputs to LLMs may not be as effective in handling this multimodal task. Furthermore, compared to the baseline methods, our proposed instruction template and retrieval module could further enhance the performance of MLLMs, surpassing previous methods and achieving the SOTA performance on MMSD2.0. This demonstrates the effectiveness of our method. Compared to the previous SOTA method on MMSD2.0, Multi-view CLIP (Qin et al., 2023), which exhibits a performance of P of 80.33, R of 88.24, and F1 of 84.10, our method demonstrates a more balanced performance across P, R, F1, approximately achieving a uniform score of 86.3 in each category.

4.5 Out-of-Domain Results

As shown in Table 4, models trained on the MMSD dataset perform poorly on the OOD dataset. Only Multi-view CLIP (Qin et al., 2023) shows relatively better performance, yet still experiences a decline of over 12 points of accuracy compared to the in-domain situation. This also indicates that MMSD indeed causes models to focus more on domain-specific data features, or even solely on textual modality features, significantly damaging the generalization of models.

Compared to MMSD, models trained on the MMSD2.0 dataset exhibit better cross-modality dependence and generalization, as shown in Table 4, which is consistent with Section 4.4. But compared to the in-domain situations, there is still a noticeable performance decline. It has been observed that the models that perform well on MMSD2.0 also tend to show decent performance on RedEval. However, models that perform well on MMSD but poorly on MMSD2.0 like DynRT-Net (Tian et al., 2023) exhibit poor performance in the OOD situation. For the OOD decline performance, we Compared to previous methods, both LLM-based methods like LLaMA2-7B and MLLM-based method LLaVA show excellent performance, even surpassing the previous SOTA methods like Multi-view CLIP. This suggests that LLMs and MLLMs indeed have better generalization capabilities compared to other models. Moreover, our method still achieves the best accuracy and F1 scores on the OOD dataset, which further demonstrates the effectiveness of our approach under the OOD condition.

Model	MMSD2.0			
	Acc. (%)	P (%)	R (%)	F1 (%)
Demon Quality	79.20	79.36	78.90	78.90
ChatGLM2-6B$^{\nu}$	80.08	80.52	81.04	80.04
ChatGLM2-6B$^{\nu}$w/RM	**82.94**	**82.62**	**83.12**	**82.76**
LLaMA2-7B$^{\nu}$	84.68	84.40	84.94	84.53
LLaMA2-7B$^{\nu}$w/RM	**85.97**	**85.64**	**85.85**	**85.74**

Table 5: The performance of retrieved demonstration quality and LLM-based methods on the MMSD2.0 dataset. The best results are highlighted in bold. ν denotes as the same in Table 3. RM denotes the retrieval module in Figure 2.

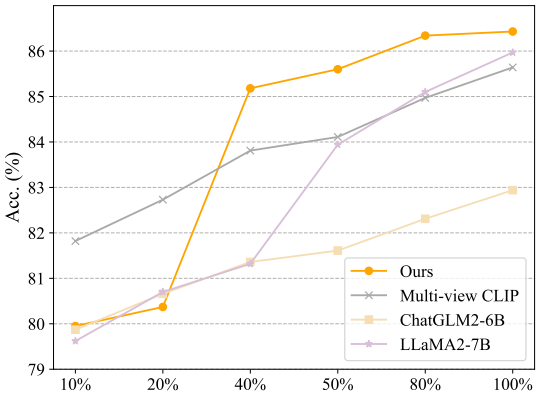

Figure 3: Low-Resource Performance on MMSD2.0.

For the observed performance degradation on RedEval, we speculate that it may be due to the source difference of multimodal samples between the RedEval dataset and the in-domain datasets. As described in Section 4.1, the datasets come from different social media platforms at different times, resulting in inherent differences in data distribution. From the statistics in Table 2, the average word count of RedEval is also lower than that of the in-domain dataset, indirectly reflecting the discrepancy between the datasets. Such differences in datasets leading to performance discrepancies further underscore the necessity of validating the generalization of models under OOD scenarios.

5 Analysis

5.1 The Effectiveness of Retrieval Module

In this section, we further analyze the effectiveness of our proposed retrieval module. Despite the results of the MLLM-based method in the main results Table 3, we further analyze the demonstration quality and the effectiveness of the retrieval module on LLM-based methods. Given that the performance of LLM-based methods on the MMSD dataset may reach the performance ceiling, we conduct experiments only on the MMSD2.0 dataset. We retrieve demonstrations for the given image-text pairs, requiring the LLMs to predict the corresponding label of the current sample based on the given demonstration. We keep adopting the format of image captions for the image information inputs.

We first evaluate the quality of our retrieved demonstrations. Generally, when the given image-text pair and its corresponding demonstration belong to the same category, the performance of models is better (Liu et al., 2022b). Therefore, we calculate the evaluation metrics by comparing the labels of the demonstrations with the corresponding image-text pairs. As shown in Table 5, with only the retrieval module, the performance of the retrieved demonstrations MMSD2.0 even surpasses most of the current multimodal methods, which are fine-tuned on the full training set. This indicates that our proposed retrieval module is highly effective in retrieving candidate demonstrations.

For the LLM-based methods, after incorporating our proposed retrieval module, they both achieve a further improvement, as shown in Table 5. Moreover, the performance of the LLaMA2-7B model even surpasses the previous SOTA method Multi-view CLIP (Qin et al., 2023). This indicates that even a purely textual LLM can achieve excellent performance in multimodal sarcasm detection by converting images into captions and applying proper prompting and instructions. This also reveals that LLMs not only have immense applicability in the field of natural language processing but also possess significant potential in the realm of multimodal tasks.

5.2 Low-Resource Scenario

Following Qin et al. (2023), we explore the effectiveness of our method and LLM-based methods in low-resource scenarios of MMSD2.0. Specifically, we compare the performance of our method, LLM-based methods with the retrieval module, and the previous SOTA method on MMSD2.0, Multi-view CLIP (Qin et al., 2023) in low-resource scenarios.

As shown in Figure 3, our MLLM-based method does not outperform Multi-view CLIP until the data proportion is above 40%. This may be because our method is based on a large scale of parameters of LLM and could not be trained sufficiently with the very limited data, leading to relatively poor perfor-

mance. With the continuous increase in data proportion in low-resource scenarios, we can observe that our method significantly outperforms Multi-view CLIP by a large margin after 40%. This indicates that with the continuous improvement in data scale, the performance based on the large language models can also be greatly enhanced.

For the LLM-based methods, we observe that LLaMA2-7B does not outperform Multi-view CLIP with a very limited amount of low-resource data like 50%. Beyond 50%, the performance of LLaMA2-7B gradually approaches and surpasses Multi-view CLIP. In contrast, ChatGLM2-6B consistently performs at a lower performance level, highlighting the importance of the choice of base models. The performance trends of LLM methods, as shown in Figure 3, are consistent with our method. This indicates that these LLM-based methods all have limitations with very low resources but their performance increases to a higher level once the data scale reaches a certain scale. This reveals that for LLMs, choosing an appropriate data scale is crucially important for performance.

6 Conclusion

In this paper, we focus on the problems of insufficient OOD generalization and inadequate utilization of cross-modal features in current multimodal sarcasm detection models. To build a more reliable model, we propose a generative multimodal sarcasm detection model consisting of an instruction template and a demonstration retrieval module based on the powerful multimodal large language model. Moreover, to assess the generalization of current multimodal sarcasm detection models, we also propose a new OOD test set, RedEval. Experimental results demonstrate that our method is effective and outperforms previous baselines by a large margin, achieving the SOTA performance on both in-domain MMSD2.0 and out-of-domain RedEval datasets.

Limitations

Our method is constrained by the foundational performance of LLM, the visual encoder, and the adapter themselves. Additionally, for the pure LLM-based methods, the quality of the image captions used as visual information input also limits the final performance of the model.

Ethics Statement

We affirm that our work here does not exacerbate the biases already inherent in the large language models. The dataset we crawled is sourced from the official public interfaces of Reddit, which met the requirements. The dataset is only for academic research. The quality of our dataset was confirmed by graduate students at Chinese universities who were paid properly.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC) No. U23B2052, Program for Youth Innovative Research Team of BUPT No. 2023QNTD02.

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. Multi-modal 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*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

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*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

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 pre-training 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 pre-training for unified vision-language understanding and generation. In *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. Multi-modal 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, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

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

Hui Liu, Wenya Wang, and Haoliang Li. 2022a. 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.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. 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*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, 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, Lillian Lee, et al. 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 multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10834–10845, Toronto, Canada. Association for Computational Linguistics.

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. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

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.

Yi Tay, Anh Tuan Luu, 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 (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, pages 162–169.

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.

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.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

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*.

A Instruction Template for LLMs

- **Pure LLM:** "Please select the sarcasm label of **'<sample text>'** from {0,1}."
- **LLM with Image Captions:** "Please select the sarcasm label of **'<sample text ### sample image caption>'** from {0,1}."
- **LLM with Demonstrations:** "Here is a demonstration: **'<demonstration text ### demonstration image caption>'**, label: '**<demonstration label>**'. Based on the above demonstration, please select the sarcasm label of **'<sample text ### sample image caption>'** from {0,1}."