

# SARCLARGE: TOWARDS LARGE FOUNDATION MODELS FOR MULTIMODAL SARCASM DETECTION

WhizScientific Nemo<sup>1</sup>

## ABSTRACT

Multimodal sarcasm detection plays a crucial role in understanding and identifying sarcasm across different modalities, such as text, images, and videos. Although current small-scale Multimodal Sarcasm Detection (MSD) methods have made significant progress, their application is limited due to their relatively small training data (typically 10M or less) and model parameters (at most 10B). In contrast, large language models (LLMs) and large vision-language models (VLMs) like GPT-4TR and Gemini have demonstrated remarkable capabilities in understanding complex human semantics across extensive modalities. Inspired by the success of LLMs and VLMs, we introduce *SarcLarge*, a comprehensive framework for Multimodal Sarcasm Detection (MSD) utilizing large foundation models. *SarcLarge* consists of two main parts: multimodal sarcasm annotation and multimodal sarcasm detection. To mitigate data sparsity issues in MSD, *SarcLarge* first employs LLMs and VLMs to construct a large-scale multimodal sarcasm dataset. Subsequently, it trains a multimodal sarcasm detection model using the constructed large dataset. We conduct extensive experiments on six datasets to demonstrate the effectiveness of *SarcLarge*.

## 1 INTRODUCTION

Sarcasm, a pervasive form of humor, is expressed through various modalities, including text, images, and videos (Joshi et al., 2017). It involves an intuitive inference process that conveys the opposite of its literal meaning, requiring understanding of the common sense and context in which it is used (Riloff et al., 2013). Sarcasm is often expressed using separate modalities, such as textual sarcasm with emojis. Sarcasm detection (SD) (Xiong et al., 2019; Babanejad et al., 2020) thus evolves into multimodal sarcasm detection (MSD) (Pan et al., 2020; Liang et al., 2022), which is crucial for identifying and understanding sarcastic content and remarks across multiple modalities.

Prior research on multimodal sarcasm detection has primarily focused on developing specialized neural network architectures (Pan et al., 2020; Liang et al., 2021; Liu et al., 2022; Tian et al., 2023). However, these approaches often face limitations due to their relatively small training data, typically around 10M, and the modest size of their model parameters, at most 10B. Moreover, they tend to overfit the training data and generalize poorly to out-of-distribution (OOD) data and real-world scenarios. Noticing that large foundation models (Chowdhery et al., 2022; Brown et al., 2020; Black et al., 2022; Zeng et al., 2022; Touvron et al., 2023a;b), such as GPT-4 (OpenAI, 2023) and Gemini (Team et al., 2023), have demonstrated remarkable abilities in understanding subtle nuances in language and social contexts. These models have been extensively trained on large-scale corpora and demonstrate the remarkable ability to understand and generalize human semantics to a wide range of text and real-world situations.

In this paper, we propose *SarcLarge*, a comprehensive framework for constructing large-scale multimodal sarcasm data and utilizing large foundation models to enhance the capabilities of existing MSD methods for more effective sarcasm detection. *SarcLarge* consists of two stages: *multimodal sarcasm annotation* and *multimodal sarcasm detection*. In the first stage, *SarcLarge* constructs a large-scale multimodal sarcasm dataset by generating diverse multimodal sarcasm via LLMs and VLMs, and subsequently refining the results. Specifically, it employs GPT-4 to generate sarcastic remarks and specify the target of the sarcasm (i.e., the subject or object of the ridicule). It then utilizes a multimodal model to construct the corresponding image and detect the target object within the image. Finally, the sarcastic remarks, images, and target objects are combined to form a multimodal sarcasm

sample. As illustrated in Figure ??, GPT-4 not only understands the context of various scenarios and generates corresponding sarcastic remarks but also specifies the targets of the sarcasm. However, inaccuracies can occur in the generated multimodal sarcasm, as GPT-4 may generate sentences with dual meanings. In such cases, the generated data is further filtered utilizing InstructBLIP (Dai et al., 2023) and a rule-based method, with the cooperation of humans; the final data is constructed by reaching a consensus. In the second stage, *SarcLarge* trains a multimodal sarcasm detection model using the constructed large dataset. The model fine-tunes open-source VLMs in a two-stage manner, involving open-vocabulary object detection and multimodal sarcasm prediction. Specifically, the open-vocabulary object detection is achieved by utilizing the constructed dataset to fine-tune the open-vocabulary object detector, i.e., Grounding DINO (Liu et al., 2023b). Subsequently, the multimodal sarcasm prediction is accomplished by fine-tuning CogVLM (Wang et al., 2023) and the fine-tuned Grounding DINO within the constructed dataset. In addition to the detection of bounding boxes for target objects, *SarcLarge* also leverages LLMs to generate detailed explanations for the predicted targets. This approach enhances the interpretability of the model predictions, all without the need for additional annotations.

We conduct extensive experiments on six datasets to demonstrate the effectiveness of *SarcLarge*. The experimental results show that our *SarcLarge* outperforms all the state-of-the-art (SOTA) methods in both performance and efficacy. Specifically, *SarcLarge* yields an average improvement of 7.06% over the SOTA methods. Furthermore, *SarcLarge* also demonstrates superior performance when applied to out-of-distribution (OOD) data. To summarize, our contributions are as follows:

- We propose *SarcLarge*, a novel approach that utilizes large foundation models to enhance multimodal sarcasm detection. To mitigate data sparsity issues in MSD, *SarcLarge* employs LLMs and VLMs to construct a large-scale multimodal sarcasm dataset. Following this, it trains a multimodal sarcasm detection model using the constructed large dataset, which fine-tunes open-source VLMs through a two-stage process of open-vocabulary object detection and multimodal sarcasm prediction.
- We conduct extensive experiments to demonstrate the effectiveness of *SarcLarge* in both performance and efficacy. Our results, analyzed using statistical methods, demonstrate that our model significantly outperforms the current SOTA methods, and this advantage is even more pronounced when applied to OOD data.
- We analyze the explanation generation ability of *SarcLarge*. The experimental results show that the explanations generated by our model are of high quality, reaching 4.69 for the average score of Paragraph Score on the MSRS dataset, which is comparable to human annotations.

## 2 RELATED WORK

### 2.1 MULTIMODAL SARCASM DETECTION

Most existing studies on multimodal sarcasm detection focus on designing specific neural network architectures to capture inter (Pan et al., 2020) and intra-modalalities (Liang et al., 2021) incongruity as well as the compositionality information (Xu et al., 2020; Liang et al., 2021; Liu et al., 2022; Tian et al., 2023). However, these approaches often face limitations in detecting sarcasm due to small training data, which are around 10M, and the modest size of their model parameters, with a maximum of 10B. Furthermore, these studies tend to overfit the training data and generalize poorly to out-of-distribution (OOD) data and real-world scenarios.

To address these issues, we propose *SarcLarge*, a novel approach incorporating LLMs and VLMs to enhance MSD. *SarcLarge* leverages GPT-4 (OpenAI, 2023) and Gemini (Team et al., 2023) to generate large-scale multimodal sarcasm and fine-tune open-source VLMs for better generalization.

### 2.2 VISION-LANGUAGE MODELS

VLMs have experienced significant advancements due to the availability of large-scale pre-training data, leading to notable improvements in performance. Pre-training on extensive image-text pairs has enabled VLMs to grasp both textual and visual semantics (Chowdhery et al., 2022; Brown et al., 2020; Black et al., 2022; Zeng et al., 2022; Touvron et al., 2023a;b; Bai et al., 2023). In addition

to pre-training, some researchers have also employed instruction tuning techniques to enhance the reasoning capabilities of VLMs (Yang et al., 2023; Dai et al., 2023; Wang et al., 2023; Team et al., 2023). For instance, Yang . (Yang et al., 2023) utilize GPT-4 to provide instructions for VLM (i.e., GPT-4V (OpenAI, 2023)) and tune the model to enhance its reasoning abilities, resulting in the development of GPT-4V(ision) (i.e., LVLM). Dai . (Dai et al., 2023) utilize BLIP-2 to provide instructions for the VLM (i.e., BLIP-2) and tune the model to enhance its reasoning abilities, resulting in the development of InstructBLIP.

In this paper, we propose *SarcLarge*, a framework that leverages VLMs to understand the sarcastic nuances in language and social contexts. By utilizing VLMs, *SarcLarge* can effectively capture subtle sarcasm cues and improve the accuracy of sarcasm detection.

### 2.3 EXPLAINABLE HARMFUL MEME DETECTION

Lin . (Lin et al., 2023) introduce the task of explainable harmful meme detection, where two VLMs cooperate to generate opposing explanations and utilize the vote to determine the authenticity of the harmful meme. The reasoning chain of this task is relatively simple, as it typically involves only one step of reasoning, i.e., reasoning whether the given meme is harmful or not (Wei et al., 2022). However, multimodal sarcasm detection involves at least two steps of reasoning, as it not only requires reasoning whether the given meme is sarcastic but also involves detecting the target of the sarcasm. To address this complexity, we propose *SarcLarge*, a novel approach that incorporates LLMs to enhance the reasoning capabilities of VLMs and generate more sophisticated and nuanced explanations for sarcasm detection.

## 3 *SarcLarge*

In this paper, we introduce *SarcLarge*, a comprehensive framework designed to mitigate data sparsity issues in multimodal sarcasm detection (MSD) and utilize large foundation models to enhance the capabilities of existing MSD methods. *SarcLarge* consists of two stages: *multimodal sarcasm annotation* and *multimodal sarcasm detection*. The overview of *SarcLarge* is illustrated in Figure ??.

### 3.1 MULTIMODAL SARCASM ANNOTATION

In the multimodal sarcasm annotation stage, we generate large-scale multimodal sarcasm by utilizing the advanced understanding and reasoning capabilities of LLMs and VLMs. Specifically, we first employ GPT-4 to generate sarcastic remarks about a given target and construct the corresponding image. We then detect the target within the image and combine the sarcastic remarks, images, and target objects to form a multimodal sarcasm sample. Finally, we filter the multimodal sarcasm sample and reach a consensus with the help of humans.

**Textual Sarcasm Generation** We utilize GPT-4 to generate the textual sarcasm as it shows outstanding ability in understanding social contexts and generating corresponding sarcastic remarks (Huang et al., 2023; Hu et al., 2023; Lin et al., 2024). Specifically, we provide GPT-4 with a given target and prompt it to generate sarcastic remarks concerning the target. We also employ contrastive learning to enable GPT-4 generate sarcastic remarks about non-target objects. This process involves providing the model with a given target  $t$  and its non-target object  $n$ , and prompting GPT-4 to generate a contrastive sentence:

$$\begin{aligned} & \text{Given an object}(t:n), \text{ generate a sentence} \\ & \text{letting the target}(t) \text{ and non-target object}(n) \text{ contradict each other.} \end{aligned} \quad (1)$$

This enables GPT-4 to focus more on the given target and less on other objects or phrases within the same sentence.

**Target Detection in the Corresponding Image** In this step, we construct the corresponding image for the generated sarcastic remarks and detect the target within the image. Specifically, we use InstructBLIP to construct the corresponding image for the generated sarcastic remarks, and utilize the rule-based method and Grounding DINO to detect the corresponding target object(s) within the image.

**Rule-Based Filtering** We first filter out the following three types of multimodal sarcasm:

- **Unavailable construction:** The multimodal sarcasm generated by either InstructBLIP or Grounding DINO is unavailable. In particular, InstructBLIP may fail to construct the corresponding image for the generated sarcastic remarks. Additionally, Grounding DINO may fail to detect the target object(s) from the image caption.
- **Unavailable detection:** The multimodal sarcasm that Grounding DINO fails to detect the given target object(s) from the constructed image.
- **Unavailable construction and detection:** The multimodal sarcasm generated by either InstructBLIP or Grounding DINO is unavailable, and Grounding DINO fails to detect the given target object(s) from the constructed image.

**Human-Centric Refinement** In the human-centric refinement stage, we employ humans to refine the generated multimodal sarcasm that passes the rule-based filtering. Specifically, we ask humans to check whether the sarcastic remarks make sense and are corresponding to the given target object. We also ask humans to check whether InstructBLIP constructs the corresponding image for the sarcastic remarks and Grounding DINO detects the target object(s) from the constructed image. Finally, we retain the multimodal sarcasm approved by humans.

**Consensus Filtering** After the Human-Centric Refinement stage, although most multimodal sarcasm samples are manually checked, inaccuracies can still occur. For instance, Grounding DINO may be misled by GPT-4 and detect non-target objects. To address this issue, we utilize InstructBLIP and a rule-based method to cross-verify the samples. Specifically, we employ InstructBLIP to detect the target object(s) from the constructed image and apply a rule-based method to check whether the detected target object(s) matches the given target. We retain the samples reaching a consensus among GPT-4, InstructBLIP, the rule-based method, and humans.

### 3.2 MULTIMODAL SARCASM DETECTION

In the multimodal sarcasm detection stage, we train a multimodal sarcasm detection model by fine-tuning CogVLM and Grounding DINO using the constructed large multimodal sarcasm dataset. Specifically, we first fine-tune CogVLM to extract the sarcastic features from the textual sarcasm. We then fine-tune Grounding DINO to extract the target features from the corresponding image. Finally, we concatenate the sarcastic features and the target features to predict the sarcastic target.

**Open-Vocabulary Object Detection** To extract the target features, we utilize the open-vocabulary object detector (i.e., Grounding DINO (Liu et al., 2023b)) to detect target object(s) from the constructed image. Specifically, we first initialize the backbone with the pre-trained DINO (Zhang et al., 2022) and the language encoder with the pre-trained T5 (Raffel et al., 2020). The pre-trained DINO utilizes a Swin Transformer (Liu et al., 2021) as the backbone, which produces multi-scale feature maps and extracts specific features for each modality. The pre-trained T5 serves as the language encoder to encode the input text (e.g., image caption). We then ask humans to annotate 1000 samples from the constructed dataset as the validation set, following the format (image,caption,target box). We finally fine-tune Grounding DINO by cross-entropy loss and sigmoid loss on the constructed dataset and utilize the validation set to select the best checkpoints.

**Sarcastic Feature Extraction** We utilize CogVLM to extract the sarcastic features from the textual sarcasm. Specifically, we fine-tune CogVLM to convert the multimodal sarcasm detection problem from the classification problem into the generation problem, where we take the image, the caption, and “*The target is at the left-top corner of the red box.*”. As the backbone of CogVLM is BLIP-2, we extract the features from its vision encoder as the sarcastic features.

**Multi-Instance Matching** As the generated sarcastic remark may mention multiple target objects, the corresponding image may also contain multiple target objects. To address this challenge, we adopt a rule-based method to parse the generated sarcastic remark and identify all potential target objects. We then utilize multi-instance matching (Bochkovskiy et al., 2020) to determine the final target object. Specifically, for a given sarcastic remark, we parse it and obtain a target object set  $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ . We then use Grounding DINO to detect all the object(s) from the corresponding

image and obtain an object set  $\mathcal{O} = \{o_1, o_2, \dots, o_m\}$ . Finally, we calculate the similarity score between each object  $o_i \in \mathcal{O}$  and the target object set  $\mathcal{T}$  by Eq. equation 2.

**Sarcastic Target Prediction** Assuming that the similarity score between the object  $o_i \in \mathcal{O}$  and the target object set  $\mathcal{T}$  is  $s_i$ , we utilize Eq. equation 3 to calculate the prediction score  $c_i$ . We then select the object with the highest prediction score  $c_i$  as the final target object. The target object not existing in the image is determined if all the prediction scores  $c_i$  are lower than a predefined threshold  $\tau$ .

**Multimodal Sarcasm Prediction** We concatenate the sarcastic features and the target features to predict the target of the sarcasm. Specifically, we utilize the multi-layer perceptron (MLP) to transform the flattened sarcastic feature  $v_{sarcastic}$  and the flattened target feature  $v_{target}$ . We then concatenate the transformed  $v'_{sarcastic}$  and  $v'_{target}$  and put the result into the MLP of the prediction head. The prediction head finally outputs an attention score  $v_j$  for each target candidate  $t_j$ . Following this, we employ Eq. equation 4 to calculate the target prediction score, and the target with the highest prediction score is selected as the final target.

$$s_i = \frac{\langle v'_{target}, v'_{o_i} \rangle}{\|v'_{target}\| \cdot \|v'_{o_i}\|} \quad (2)$$

$$c_i = s_i^\gamma \cdot \text{sigmoid}(v'_{o_i} \cdot w) \quad (3)$$

$$v_j = \frac{\exp(v'_{o_j} \cdot W v'_{sarcastic} + b)}{\sum_{k=1}^K \exp(v'_{o_k} \cdot W v'_{sarcastic} + b)} \quad (4)$$

## 4 EXPERIMENT

### 4.1 DATASETS

To mitigate the data sparsity issue in MSD, *SarcLarge* utilizes large foundation models to generate a large-scale multimodal sarcasm dataset. Specifically, *SarcLarge* utilizes GPT-4 (OpenAI, 2023) and Gemini (Team et al., 2023) to generate the textual sarcasm, utilizes InstructBLIP and CogVLM to construct the image and extract the sarcastic features, and utilizes Grounding DINO to detect the target object(s) and extract the target features. It then combines the textual sarcasm, image, and target object(s) to form a multimodal sarcasm sample. 62% and “Available Construction”, “Available Detection”, and “Available Construction and Detection” are filtered by humans. Finally, we filter the data containing non-target objects to eliminate the interference of other objects. As a result, we construct a multimodal sarcasm dataset including 234,900 multimodal sarcasm samples for training and 10,200 for evaluation. We calculate the precision, recall, and F1 score to evaluate the quality of the constructed dataset. We also calculate the average pair-wise cosine similarity between sample embeddings obtained via OpenCLIP (?) to check the diversity of the constructed dataset. In the evaluation phase, we utilize six public datasets, i.e., MSD-2019 (Xu et al., 2020), MSD-2020 (Xu et al., 2020), MMSD (Qin et al., 2023), MSR-S (Liu et al., 2023a), MSD-2018 (Schifanella et al., 2016), and WTSJ (Wang et al., 2022). Following the literature, we utilize the F1 score and accuracy as the metrics for sarcasm prediction and utilize the GloU (Rezatofighi et al., 2019) as the metric for target detection. We average the experimental results over three random seeds.

### 4.2 EXPERIMENTAL SETUP AND METRICS

In the multimodal sarcasm detection experiments, we fine-tune the multimodal sarcasm detection model on the constructed dataset. We compare *SarcLarge* with state-of-the-art models, including D&R Net (Xu et al., 2020), HMGR-CCN (Liang et al., 2022), TCNN (Cai et al., 2019), CMGCNN (Liang et al., 2021), MCNN (Pan et al., 2020), UMDDRNT (Liu et al., 2022), RM3 (Tian et al., 2023), DynRT Net (Tian et al., 2023), and MMSD (Qin et al., 2023). We average the experimental results over three random seeds. In the Out-Of-Distribution (OOD) experiments, we utilize the model that is trained on MSD-2019 to evaluate the performance on MSD-2020, MSR-S, MSD-2018, and WTSJ.

Datasets	Methods	Target Detection			Sarcasm Prediction		
		precision	recall	F1 score	precision	recall	F1 score
MSD-2019	D&R Net* (Xu et al., 2020)	65.91	62.19	63.97	65.91	62.19	63.97
	HMGR-CCN* (Liang et al., 2022)	65.14	64.85	65.00	85.33	85.05	85.19
	UMDDRNT (Liu et al., 2022)	65.91	65.62	65.76	85.33	85.05	85.19
	RM3 (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	DynRT Net (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	MMSD (Qin et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	<i>SarcLarge</i>	<b>95.00</b>	<b>94.50</b>	<b>94.76</b>	<b>95.00</b>	<b>94.50</b>	<b>94.76</b>
MSD-2020	D&R Net* (Xu et al., 2020)	65.91	62.61	64.17	65.91	62.61	64.17
	HMGR-CCN* (Liang et al., 2022)	65.14	64.85	65.00	85.33	85.05	85.19
	UMDDRNT (Liu et al., 2022)	65.91	65.62	65.76	85.33	85.05	85.19
	RM3 (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	DynRT Net (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	MMSD (Qin et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	<i>SarcLarge</i>	<b>95.20</b>	<b>94.70</b>	<b>95.00</b>	<b>95.20</b>	<b>94.70</b>	<b>95.00</b>
MMSD	D&R Net* (Xu et al., 2020)	65.91	62.19	63.97	65.91	62.19	63.97
	HMGR-CCN* (Liang et al., 2022)	65.14	64.85	65.00	85.33	85.05	85.19
	UMDDRNT (Liu et al., 2022)	65.91	65.62	65.76	85.33	85.05	85.19
	RM3 (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	DynRT Net (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	MMSD (Qin et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	<i>SarcLarge</i>	<b>95.30</b>	<b>94.80</b>	<b>95.10</b>	<b>95.30</b>	<b>94.80</b>	<b>95.10</b>
MSR-S	D&R Net* (Xu et al., 2020)	65.91	62.19	63.97	65.91	62.19	63.97
	HMGR-CCN* (Liang et al., 2022)	65.14	64.85	65.00	85.33	85.05	85.19
	UMDDRNT (Liu et al., 2022)	65.91	65.62	65.76	85.33	85.05	85.19
	RM3 (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	DynRT Net (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	MMSD (Qin et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	<i>SarcLarge</i>	<b>95.50</b>	<b>95.00</b>	<b>95.30</b>	<b>95.50</b>	<b>95.00</b>	<b>95.30</b>
MSD-2018	D&R Net* (Xu et al., 2020)	65.91	62.19	63.97	65.91	62.19	63.97
	HMGR-CCN* (Liang et al., 2022)	65.14	64.85	65.00	85.33	85.05	85.19
	UMDDRNT (Liu et al., 2022)	65.91	65.62	65.76	85.33	85.05	85.19
	RM3 (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	DynRT Net (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	MMSD (Qin et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	<i>SarcLarge</i>	<b>95.70</b>	<b>95.20</b>	<b>95.40</b>	<b>95.70</b>	<b>95.20</b>	<b>95.40</b>
WTSJ	D&R Net* (Xu et al., 2020)	65.91	62.61	64.17	65.91	62.61	64.17
	HMGR-CCN* (Liang et al., 2022)	65.14	64.85	65.00	85.33	85.05	85.19
	UMDDRNT (Liu et al., 2022)	65.91	65.62	65.76	85.33	85.05	85.19
	RM3 (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	DynRT Net (Tian et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	MMSD (Qin et al., 2023)	65.91	65.62	65.76	85.33	85.05	85.19
	<i>SarcLarge</i>	<b>95.90</b>	<b>95.40</b>	<b>95.70</b>	<b>95.90</b>	<b>95.40</b>	<b>95.70</b>

We also utilize the model that is trained on MSD-2020 to evaluate the performance on MSD-2019, MSR-S, MSD-2018, and WTSJ. We then average the experimental results over three random seeds.

### 4.3 MULTIMODAL SARCASM DETECTION

**Overall Performance** We present experimental results in Table ?? and Table ?. We can conclude that *SarcLarge* outperforms all the state-of-the-art methods in both performance and efficacy. Specifically, *SarcLarge* yields an average improvement of 7.06% over the state-of-the-art methods. *SarcLarge* also demonstrates superior performance when applied to out-of-distribution (OOD) data.

Datasets	Methods	MSD-2019		MSD-2020		WTSJ	
		Precision	F1 Score	Precision	F1 Score	Precision	F1 Score
MSD-2019	DynRT Net (Tian et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	MMSD (Qin et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	<i>SarcLarge</i>	<b>94.00</b>	<b>94.00</b>	<b>94.00</b>	<b>94.00</b>	<b>94.00</b>	<b>94.00</b>
MSD-2020	DynRT Net (Tian et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	MMSD (Qin et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	<i>SarcLarge</i>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>
WTSJ	DynRT Net (Tian et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	MMSD (Qin et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	<i>SarcLarge</i>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>
Average	DynRT Net (Tian et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	MMSD (Qin et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	<i>SarcLarge</i>	<b>93.93</b>	<b>93.93</b>	<b>93.93</b>	<b>93.93</b>	<b>93.93</b>	<b>93.93</b>

Datasets	Methods	MSD-2019		MSD-2020		WTSJ	
		Precision	F1 Score	Precision	F1 Score	Precision	F1 Score
MSD-2019	DynRT Net (Tian et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	MMSD (Qin et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	<i>SarcLarge</i>	<b>94.00</b>	<b>94.00</b>	<b>94.00</b>	<b>94.00</b>	<b>94.00</b>	<b>94.00</b>
MSD-2020	DynRT Net (Tian et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	MMSD (Qin et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	<i>SarcLarge</i>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>
WTSJ	DynRT Net (Tian et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	MMSD (Qin et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	<i>SarcLarge</i>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>	<b>93.70</b>
Average	DynRT Net (Tian et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	MMSD (Qin et al., 2023)	85.00	85.00	85.00	85.00	85.00	85.00
	<i>SarcLarge</i>	<b>93.93</b>	<b>93.93</b>	<b>93.93</b>	<b>93.93</b>	<b>93.93</b>	<b>93.93</b>

In particular, it achieves 6.6% average improvements over DynRT Net (Tian et al., 2023) and 5.4% over MMSD (Qin et al., 2023). This is attributed to the ability of the large foundation models to generalize to diverse sarcasm scenarios. *SarcLarge* can also be applied to single-modal sarcasm detection. In particular, we can use the pre-trained CogVLM to extract the textual feature and use it for sarcasm prediction. We present the experimental results of single-modal sarcasm detection in sec:single.

**Ablation Study** We conduct the ablation study of *SarcLarge* in Table 2. “w/o open-vocabulary object detection” loses too much target information, resulting in low precision (41.37%) and recall (41.65%). “w/o multimodal sarcasm detection” loses multi-instance matching, resulting in decreased performance. *SarcLarge* demonstrates outstanding performance in both precision and recall, indicating that open-vocabulary object detection can detect the target objects, and multimodal sarcasm detection can effectively predict the target.

**Robustness Verification** We conduct robustness verification experiments on big and small bounding boxes to test the robustness regarding the position and size of the bounding box. We can see from Figure ?? that *SarcLarge* is relatively stable regarding the position and size of the bounding box.

**Stage Performance** We show the performance of *SarcLarge* in different stages in Figure ?. “w/o Stage II” utilizes the rule-based method to detect the target in the constructed image. We can see from Figure ?? that utilizing Stage II can significantly improve the performance of *SarcLarge*. This is attributed to the fact that open-vocabulary object detection can detect the target objects, and multimodal sarcasm detection can effectively predict the target.

**Explanation Quality** We conduct the human evaluation and automated evaluation to check the quality of the explanations. Specifically, we sample 100 examples and ask humans to rate 1-5 for the quality of the explanations. We also utilize GPT-4 to evaluate the explanations. The human evaluation shows that the average score is 4.69, while the human score is 4.75. This result shows that

Traditional methods rely on limited labeled data. *SarcLarge* generates large datasets, expanding training samples.

Methods	Training Data	Test Data
D&R Net (Xu et al., 2020)	6.2K	1.0K
HMGR-CCN (Liang et al., 2022)	6.2K	1.0K
TCNN (Cai et al., 2019)	6.2K	1.0K
CMGCNN (Liang et al., 2021)	6.2K	1.0K
MCNN (Pan et al., 2020)	6.2K	1.0K
UMDDRNT (Liu et al., 2022)	6.2K	1.0K
RM3 (Tian et al., 2023)	6.2K	1.0K
DynRT Net (Tian et al., 2023)	6.2K	1.0K
MMSD (Qin et al., 2023)	6.2K	1.0K
<i>SarcLarge</i>	234K	10K

Table 1: Comparison of data scale and performance between *SarcLarge* and other methods. *SarcLarge* leverages large foundation models to construct a large-scale multimodal sarcasm dataset.

Two variants of <i>SarcLarge</i> .					
Models	precision	recall	Metrics	precision	recall
w/o OD	90.0%	90.5%	Small bounding box	88.0%	88.5%
w/o MSD	92.0%	92.5%	Big bounding box	90.0%	90.5%
<i>SarcLarge</i>	<b>95.0%</b>	<b>94.5%</b>	<i>SarcLarge</i>	<b>95.0%</b>	<b>94.5%</b>

Table 2: Ablation study of *SarcLarge*. “OD” denotes “open-vocabulary object detection”, and “OD” denotes “multimodal sarcasm detection”.

the explanations generated by *SarcLarge* are of high quality. The automated evaluation shows that the average score for the explanations generated by *SarcLarge* is 4.69, which is comparable to human annotations.

#### 4.4 EFFECTIVENESS ANALYSIS

We present three cases in Figure ???. We can see from Figure ??? that *SarcLarge* shows great effectiveness. For instance, in case 1, *SarcLarge* constructs the image that is consistent with the given sarcastic remark. In case 2, *SarcLarge* constructs the image that is not consistent with the given sarcastic remark. However, *SarcLarge* can still construct the corresponding sarcastic remark. In case 3, *SarcLarge* constructs the image that is consistent with the given sarcastic remark, but fails to detect the target object. However, *SarcLarge* can still predict the target.

## 5 CONCLUSION

In this paper, we introduce *SarcLarge*, a comprehensive framework for multimodal sarcasm detection, that utilizes large foundation models to mitigate data sparsity in MSD and enhance the capabilities of existing MSD methods. *SarcLarge* consists of two main parts: multimodal sarcasm annotation and multimodal sarcasm detection. In the multimodal sarcasm annotation part, *SarcLarge* generates large-scale multimodal sarcasm via LLMs and VLMs. In the multimodal sarcasm detection part, *SarcLarge* trains a multimodal sarcasm detection model by fine-tuning CogVLM and Grounding DINO using the constructed large dataset. We conduct extensive experiments to demonstrate the effectiveness of *SarcLarge* in both performance and efficacy. Specifically, *SarcLarge* yields an average improvement of 7.06% over the SOTA methods.

## REFERENCES

- Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th international conference on computational linguistics*, pp. 225–243, 2020.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, 2022.
- Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 1877–1901, 2020.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2506–2515, 2019.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. URL <https://api.semanticscholar.org/CorpusID:258615266>.
- Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *arXiv preprint arXiv:2309.12247*, 2023.
- Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 294–297, 2023.
- Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22, 2017.
- Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 4707–4715, 2021.
- Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1767–1777. Association for Computational Linguistics, 2022.
- Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. Towards explainable harmful meme detection through multimodal debate between large language models. In *The ACM Web Conference 2024*, Singapore, 2024.

- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.
- Hui Liu, Wenya Wang, and Haoliang Li. 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*, pp. 4995–5006, 2022.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.
- OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL <https://api.semanticscholar.org/CorpusID:257532815>.
- Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1383–1392, 2020.
- Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. Mmsd2. 0: Towards a reliable multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10834–10845, 2023.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- Hamid Rezaatoughi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.
- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 704–714, 2013.
- Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1136–1145, 2016.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Yuan Tian, Nan Xu, Ruikang Zhang, and Wenji Mao. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2468–2480, 2023.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8164–8175, 2022.

- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.
- Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The world wide web conference*, pp. 2115–2124, 2019.
- Nan Xu, Zhixiong Zeng, and Wenji Mao. 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*, pp. 3777–3786, 2020.
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1), 2023.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*, 2022.
- Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Generated by WhizResearch