CofiPara: A Coarse-to-fine Paradigm for Multimodal Sarcasm Target Identification with Large Multimodal Models

Hongzhan Lin[♡*], **Zixin Chen**[♣*], **Ziyang Luo**[♡], **Mingfei Cheng**[♢], **Jing Ma**[♡†], **Guang Chen**[♣†]
[♡]Hong Kong Baptist University
[♣]Beijing University of Posts and Telecommunications
[♢]Singapore Management University
{cshzlin,majing}@comp.hkbu.edu.hk, {mailboxforvicky,chenguang}@bupt.edu.cn

Abstract

Social media abounds with multimodal sarcasm, and identifying sarcasm targets is particularly challenging due to the implicit incongruity not directly evident in the text and image modalities. Current methods for Multimodal Sarcasm Target Identification (MSTI) predominantly focus on superficial indicators in an end-to-end manner, overlooking the nuanced understanding of multimodal sarcasm conveyed through both the text and image. This paper proposes a versatile MSTI framework with a coarse-to-fine paradigm, by augmenting sarcasm explainability with reasoning and pre-training knowledge. Inspired by the powerful capacity of Large Multimodal Models (LMMs) on multimodal reasoning, we first engage LMMs to generate competing rationales for coarser-grained pre-training of a small language model on multimodal sarcasm detection. We then propose fine-tuning the model for finer-grained sarcasm target identification. Our framework is thus empowered to adeptly unveil the intricate targets within multimodal sarcasm and mitigate the negative impact posed by potential noise inherently in LMMs. Experimental results demonstrate that our model far outperforms state-of-the-art MSTI methods, and markedly exhibits explainability in deciphering sarcasm as well.

1 Introduction

Sarcasm, a prevalent form of figurative language, is often used in daily communication to convey irony, typically implying the opposite of its literal meaning (Joshi et al., 2017). As an important component in deciphering sarcasm, automated Sarcasm Target Identification (STI) is crucial for Natural Language Processing (NLP) in customer service (Davidov et al., 2010), opinion mining (Riloff et al., 2013), and online harassment detection (Yin et al., 2009). Although prior research on STI has primarily centered on textual content (Joshi et al., 2018; Parameswaran et al., 2019), the surge in multimodal user-generated content has propelled the field of multimodal sarcasm target identification to the forefront of research (Wang et al., 2022), making it a significant area of study in both NLP applications and multimedia computing.

(a) (b)

Figure 1: Examples of multimodal sarcasm on **Twitter**: (**a**) "*never seen a #dlr train driver before. looks like a tough job #london*"; (**b**) "*thank god for no product placement in #ukraine #eurovision*". Boxes in green and words in red denote the visual and textual targets.

The MSTI task is to extract the entities being ridiculed (i.e., sarcasm targets) from both the text and image in multimodal sarcastic content. Previous work (Devlin et al., 2019; Bochkovskiy et al., 2020) attempted to straightforwardly integrate a BERT-based textual encoder and a CNN-based visual encoder for just modeling the sarcasm text and image, respectively. The state-of-the-art approach (Wang et al., 2022) treats MSTI merely as an end-to-end task, primarily focusing on the superficial signals evident in the surface-level text and image. However, a more thorough investigation and understanding of the underlying meanings are essential, particularly in cases where the correlation between image and text is not immediately apparent in multimodal sarcasm (Tian et al., 2023).

Comprehending and analyzing multimodal sarcasm poses a considerable challenge, because its implicit meaning demands an in-depth understanding and reasoning of commonsense knowledge. For example, as shown in Figure 1(a), a human checker

[*] Equal contribution.
[†] Corresponding authors.

needs reasonable thoughts between visual and textual sarcasm targets, to understand that the man's leisurely sitting posture in front of the control panel creates a sarcastic contrast between the idea of the train driver's job being difficult and the actual scene. Moreover, as the example shows in Figure 1(b), the sarcasm target sometimes does not appear explicitly in the text, which makes it more challenging for conventional models to cognize that the example implies that the presence of the Pepsi bottle is a form of product placement, which is often seen as a marketing tactic. We contend the challenge lies in delivering rich multimodal knowledge that consistently assists in deciphering the concealed semantics within the multimodal nature of sarcasm.

In this paper, we adhere to the following two key principles in the knowledge-augmented design of our approach: 1) LMM Reasoning: To grasp the implicit meanings intrinsic in sarcasm, we resort to the extensive prior knowledge embedded within Large Multimodal Models (LMMs) (Liu et al., 2023a; Bai et al., 2023). This design philosophy enables complex reasoning, thereby enhancing both the MSTI accuracy and explainability; 2) MSD Pre-training: Previous literature (Joshi et al., 2018) indicates that MSTI inherently appraises the presence of sarcasm targeting each entity within sarcastic content. Similar to Multimodal Sarcasm Detection (MSD) (Qin et al., 2023), which involves determining the sarcasm in multi-modalities at a holistic content level, MSTI actually engages in finer-grained sarcasm detection at the localized entity level. Considering such a close correlation between MSD and MSTI, it is assumed that insights from the coarser-grained MSD are instrumental in discerning sarcasm targets in the finer-grained MSTI. Thus we devise a cohesive framework to operate on the coarse-to-fine training paradigm, aimed at pinpointing nuanced visual and textual targets of multimodal sarcasm for MSTI by benefitting from LMM reasoning and MSD pre-training.

To these ends, we propose a novel framework with a **Co**arse-to-**fi**ne **Para**digm, **CofiPara**, by leveraging the divergent knowledge extracted from LMMs for multimodal sarcasm target identification. Specifically, we integrate text and image modalities within the coarse-to-fine training paradigm, which consists of two phases: 1) Coarser-grained Sarcasm Detection: Initially, we engage LMMs in critical thinking to generate rationales from both sarcastic and non-sarcastic perspectives. Utilizing these generated sarcasm rationales, we pre-train a smaller model to act as a rationale referee to implicitly extract sarcasm-indicative signals in the competing rationales for sarcasm prediction. This process aligns multimodal features between the sarcasm content and its underlying rationales, alleviating the negative impact of inevitable noise from LMMs through competing rationales; 2) Finer-grained Target Identification: Subsequently, we further fine-tune the smaller model pre-trained in the previous stage for multimodal sarcasm target identification. This phase enhances our model with the multimodal reasoning knowledge, acquired in the pre-training stage and the rationale in sarcastic perspective, to reveal the meanings concealed within the comprehensive multimodal information of sarcasm samples. In this manner, our CofiPara framework could be naturally output as the explanatory basis for deciphering multimodal sarcasm. Extensive experiments conducted on two public sarcasm datasets reveal that our approach far outperforms previous state-of-the-art MSTI methods, and achieves competitive results compared with MSD baselines. The experimental analysis further underscores the enhanced ability to provide superior explainability in the realm of multimodal sarcasm. Our contributions are summarized as follows in three folds:

- To the best of our knowledge, we are the first to study multimodal sarcasm from a fresh perspective on explainability in both multimodal targets and natural texts, by exploiting advanced large multimodal models. [1]
- We propose a universal MSTI framework with the novel coarse-to-fine paradigm that incorporates the multimodal sarcasm target identification and the textual explanation for deciphering the multimodal sarcasm, which enhances sarcasm explainability in conjunction with effective multimodal sarcasm detection.
- Extensive experiments confirm that our framework could yield superior performance on multimodal sarcasm target identification, and further provide informative explanations for a better understanding of multimodal sarcasm.

2 Related Work

MSD. Sarcasm detection involves discerning sentiment incongruity within a context, traditionally

[1] Our source code is available at https://github.com/Lbotirx/CofiPara.

emphasizing text modality (Xiong et al., 2019; Babanejad et al., 2020). Multimodal Sarcasm Detection (MSD), enhanced by image integration, has garnered growing research interest (Schifanella et al., 2016). Cai et al. (2019) introduced a comprehensive MSD dataset, incorporating text, image, and image attributes, alongside a hierarchical fusion model. Subsequently, a range of studies has utilized attention mechanisms to subtly blend features from different modalities (Xu et al., 2020; Pan et al., 2020; Tian et al., 2023). Another line of recent advancements has seen the introduction of graph-based methods for sarcasm detection (Liang et al., 2021, 2022; Liu et al., 2022), which excel in identifying key indicators across modalities. Qin et al. (2023) revealed spurious cues in the previous MSD dataset (Cai et al., 2019) and provided an alternative refined dataset version. Existing solutions, however, only focused on performing multimodal sarcasm classification (i.e., predicting if a sample is sarcastic) with limited explanations for its prediction. In this paper, we delve into the explainability of multimodal sarcasm from both multimodal targets and textual rationales, aiming to decipher multimodal sarcasm using more intuitive forms and assisting users in gaining a better understanding.

MSTI. Recent advancements in sarcasm analysis have seen a significant focus on Sarcasm Target Identification (STI), with notable contributions from researchers. STI aims to pinpoint the subject of mockery in sarcastic texts. Joshi et al. (2018) introduced the concept of STI and discussed its application in the 2019 ALTA shared task (Molla and Joshi, 2019), highlighting evaluation metrics like Exact Match accuracy and F1 score. Patro et al. (2019) later developed a deep learning model enhanced with socio-linguistic features for target identification, while Parameswaran et al. (2019) utilized a combination of classifiers, followed by a rule-based method for extracting textual sarcasm targets. Moreover, Wang et al. (2022) pioneered STI in multimodal contexts by integrating sequence labeling with object detection in an end-to-end manner, but only capturing the superficial signals of different modalities in sarcasm. In this work, we regard the MSD task as the predecessor pre-training phase of MSTI, to better derive the prior reasoning knowledge absorbed in the coarser-grained auxiliary task MSD to the finer-grained goal task MSTI.

LLMs and LMMs. Recently, Large Language Models (LLMs) have demonstrated exceptional versatility across various tasks. Significant advancements by leading tech companies have resulted in highly proficient, though often proprietary, LLMs (Brown et al., 2020; OpenAI, 2023; Chowdhery et al., 2022; Team et al., 2023). Meanwhile, the NLP community has seen the rise of open-source LLMs, with publicly shared model weights (Black et al., 2022; Zeng et al., 2022; Touvron et al., 2023a,b; Xu et al., 2023; Luo et al., 2023). More recently, LLMs have also been developed to adapt in processing both textual and visual data, marking a significant advancement. Recent research has focused on constructing versatile multimodal datasets (Yang et al., 2023b) from platforms like GPT-4 and GPT-4V (OpenAI, 2023), fine-tuning open-source LMMs such as LLaVA (Liu et al., 2023a), Qwen-VL (Bai et al., 2023), and other innovative projects (Dai et al., 2023; Wang et al., 2023). These LMMs have shown excellent emergent abilities in multimodal tasks. In this work, we foster divergent thinking in LMMs by employing potential sarcasm labels as prompts, which promotes a coarse-to-fine strategy for fine-tuning smaller Language Models (LMs). Combined with MSTI, this design philosophy enhances the sarcasm understanding within the universal framework, steering it towards greater sarcasm explainability.

3 Our Approach

Problem Statement. We define a multimodal sample as $M = \{I, T\}$, which consists of an image I and a text T. In the context of the coarser-grained MSD task, the label y of the sample falls into either of the categories: `sarcastic` or `non-sarcastic`. As for the finer-grained MSTI task, the label y is a tuple consisting of a textual sarcasm target y_{text}, and a visual bounding box y_{img}, where the model is tasked with pinpointing the sarcasm entity targeted within the provided text and image modalities of the *sarcastic* sample. In this paper, we focus on improving the finer-grained MSTI task by leveraging insights from the coarser-grained MSD task.

Closed to the MSD task, which establishes the presence of sarcasm in holistic semantics at the coarser level, the MSTI task inherently detects sarcasm targeting each entity of the multimodal sarcastic content to explicitly identify the specific sarcasm targets at the finer level (Joshi et al., 2018). This work is designed mainly to integrate MSD and MSTI into a versatile framework with the coarse-to-fine training paradigm, which utilizes MSD as the predecessor foundational stage to facilitate the

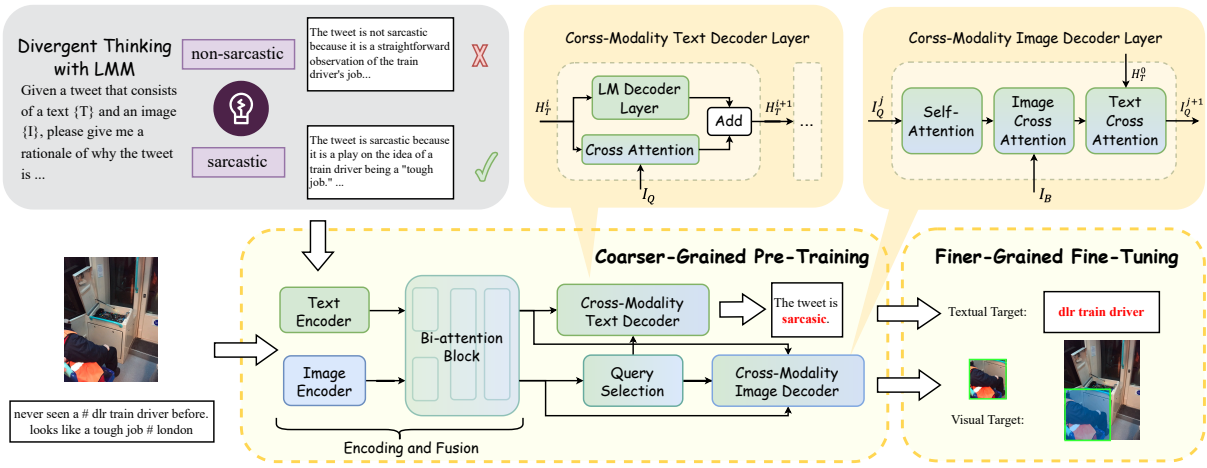


Figure 2: An overview of our framework, CofiPara, for multimodal sarcasm target identification.

subsequent MSTI process, through incorporating rationales generated from LMMs. The overview of our model is illustrated in Figure 2, which consists of: 1) Divergent Thinking with LMM (§3.1), 2) Coaser-Grained Pre-Training (§3.2), and 3) Finer-Grained Fine-Tuning (§3.3).

3.1 Divergent Thinking with LMM

Generally, LLMs can generate reasonable thoughts (Wei et al., 2022) to unveil the underlying meaning of the sarcasm. The rationales from LLMs usually express perspectives grounded with commonsense or related to certain social scenarios (Huang et al., 2023; Lin et al., 2024a), which can be used as extensive prior knowledge for smaller downstream models to facilitate decision-making. Although LLMs have shown emergent abilities in reasoning and interpreting, they still suffer from preconception bias and may generate uncredible content or even make false assertions (Hu et al., 2023; Lin et al., 2023a). Therefore, the downstream decision-maker model would have to be robust enough to alleviate the negative impact imposed by the input noisy LLM-generated rationales. In pursuit of this, we resort to the inspiration of divergent multimodal thinking with vision LLMs, i.e., LMMs, fostering our model to explore a more reliable reasoning pathway from the conflicting and noisy insights provided by LMMs.

Given an input $M = \{I, T\}$, we prompt LMMs to generate a pair of competing rationales based on the text T, the image I, and the potential sarcasm labels $* \in \{\texttt{sarcastic}, \texttt{non-sarcastic}\}$, by using a prompt template p we curated in advance. To exploit LMMs' divergent reasoning ability, we construct each sample with different potential sarcasm labels, respectively. Specifically, we design the prompt template p as follows:

"Given a tweet that consists of a text and an image, please give me a rationale of why the tweet is $\{*\}$.
tweet text: $\{T\}$
tweet image: $\{I\}$*"*

Note that the potential labels $*$ are just used to formalize two opposite standpoints for multimodal reasoning regardless of the ground-truth label. Then we can derive the competing rationales r_{pos} or r_{neg} from LMMs to support the `sarcastic` or `non-sarcastic` positions. By introducing adversarial labels, we encourage LMMs to adopt diverse perspectives, thereby providing a range of background knowledge enriched with deliberate noise. Because the rationale for the false class ideally contains more useless information than the other one for the ground truth. The contextual subtleties of sarcasm that are pivotal to rival candidate sarcasm categories, can be thus more effectively highlighted and contrasted. This allows the rest of the model to achieve the logical reasoning of the true sarcastic intent by considering it from diverse perspectives, while moderating vulnerability to the potential noise in the LMM-generated rationales.

3.2 Coarser-Grained Pre-Training

Given the close correlation between the coarser-grained multimodal sarcasm detection with the finer-grained multimodal sarcasm target identification, we advocate for initial pre-training in multimodal sarcasm detection, allowing the model to grasp the essence of sarcasm preliminarily. This foundational understanding could set the stage for a more nuanced and detailed identification of sarcasm targets in subsequent fine-tuning.

Encoding and Fusion. For an input sample $M = \{I, \hat{T}\}$ packed with the generated competing rationales, where $\hat{T} = \{T, r_{pos}, r_{neg}\}$ is the input text, we first extract textual and visual features as:

$$H_T = \mathsf{E}_T(\hat{T}), \;\; I_E = \mathsf{E}_I(I), \tag{1}$$

where $H_T \in \mathbb{R}^{m \times d}$ is the token embedding output by the text encoder $\mathsf{E}_T(\cdot)$ implemented by Transformer Encoder (Raffel et al., 2020), m is the input token length and d is the dimension of hidden states. $\mathsf{E}_I(\cdot)$ denotes the image encoder based on Vision Transformer (Liu et al., 2021), used to fetch the patch-level features of the image with n patches, which are projected into the visual features $I_E \in \mathbb{R}^{n \times d}$. Since both the text and image encoders are designed as Transformer-based, the embeddings shaped by the isomorphic encoding structure can enhance consecutive multimodal fusion during encoding (Liu et al., 2023b). Then, to align the semantics in the text and image, we adopt a bi-directional attention module based on the cross-attention mechanism. Taking the text-to-image cross-attention $CrossAttn(H_T, I_E)$ as an example, we define the query, key and value as $\{Q_T, K_I, V_I\} = \{H_T W_Q, I_E W_K, I_E W_V\}$, where $\{W_Q, W_K, W_V\} \in \mathbb{R}^{d \times d_k}$ are trainable weights. Then the calculation is as follows:

$$H_T^0 = \text{softmax}\left(\frac{Q_T K_I^\top}{\sqrt{d_k}}\right) V_I, \tag{2}$$

With Equation 2, similarly, we can calculate the image-to-text cross-attention. Combing the two cross-attention modules together, we fuse the multimodal features during encoding as follows:

$$\begin{aligned} H_T^0 &= CrossAttn(H_T, I_E), \\ I_B &= CrossAttn(I_E, H_T), \end{aligned} \tag{3}$$

where H_T^0, I_B are the attended textual and visual features, respectively. To optimize the information integration of multimodal sarcasm with competing rationales, we further develop a query selection mechanism to prioritize image region features that exhibit higher correlations with the input text. This yields:

$$\arg\max_n \left(\max_m \left(I_B {H_T^0}^\top\right)\right), \tag{4}$$

with which we obtain the index of the topmost relevant local visual features, queried by the textual features H_T^0 from the global visual features I_B. We name the selected local visual features as I_Q.

Cross-Modality Text Decoding. Based on the attended textual features H_T^0 and query-selected local visual features I_Q, we then devise a multimodal decoding strategy with textual outputs to infer sarcasm for MSD. Specifically, during decoding, we only exploit a text-to-image cross-attention module to attain the textual features attended with the visual ones:

$$H_{T_{attn}}^i = CrossAttn(H_T^i, I_Q), \tag{5}$$

where H_T^i is the textual feature input of the i^{th} LM decoder layer, and $H_{T_{attn}}^i$ is the attended textual feature output of the cross-attention $CrossAttn$. Then by adding the attended features $H_{T_{attn}}^i$ to the output of the i^{th} LM decoder LM_{dec}^i, the fused intermediate features H_T^{i+1} fed into the next LM decoder layer are:

$$H_T^{i+1} = LM_{dec}^i(H_T^i) + H_{T_{attn}}^i. \tag{6}$$

After L layers of cross-modality LM decoder, we have the final textual representations H_T^L, further decoded as the text output to clearly express whether the sample is sarcastic. Finally, we train the model f by minimizing the following loss:

$$\mathcal{L}_{text} = CE(f(I, \hat{T}), y), \tag{7}$$

where $CE(\cdot)$ denotes the cross-entropy loss between the generated label token and ground truth label y for MSD. During the coarser-grained pre-training, the model is trained to distill the essence and discard irrelevant elements from the divergent thinking of LMMs about sarcasm. Such a process could fortify our model's resilience in the subsequent fine-tuning stage for MSTI, ensuring robustness against the potential inaccuracies stemming from LMMs, leading to a more independent and refined thought of the LMM-generated rationales.

3.3 Finer-Grained Fine-Tuning

After the coarser-grained pre-training stage, our model could be resilient against the potential variation and bias in LMMs through the competing rationales, to first comprehend what constitutes sarcasm. As the goal of our approach is to identify both the textual and visual sarcasm targets for further deciphering sarcasm, we conduct the finer-grained fine-tuning stage for MSTI, which shares the same model architecture, parameters of the multimodal encoding and text decoding procedures as §3.2 but differs in the text decoding output and an additional image decoding procedure.

Different from the pre-training stage in §3.2, the sample M in MSTI is set as *sarcastic* prior due to the nature of this specific task (Wang et al., 2022). Thus the input text for a given *sarcastic* sample M is formed as $\hat{T} = \{T, r_{pos}\}$ in this stage, where we

only provide the text T and the LMM-generated rationale r_{pos} that explains why M is sarcastic. For the cross-modality text decoding, we generate the predicted textual sarcasm targets that are entities in the text T. Then the textual target loss $\hat{\mathcal{L}}_{text}$ can be computed akin to that outlined in Equation 7.

Cross-Modality Image Decoding. For visual object detection, we use a cross-modality image decoder to discern the visual sarcasm target, where the textual features H_T^0 and the global visual features I_B are used to attend to the local visual features I_Q in each Transformer decoder layer:

$$\begin{aligned} I_Q^{j'} &= SelfAttn\left(I_Q^j\right), \\ I_Q^{j''} &= CrossAttn\left(I_Q^{j'}, I_B\right), \\ I_Q^{j+1} &= CrossAttn\left(I_Q^{j''}, H_T^0\right), \end{aligned} \tag{8}$$

where $SelfAttn(\cdot)$ denotes self-attention, and I_Q^j is the input of the j^{th} Transformer decoder layer. After K layers of the image decoder, we have the final visual features I_Q^K. Afterwards, we decode I_Q^K as the image output consisting of a bounding box output and its confidence score. Following previous object detection work (Zhang et al., 2022), we use the L1 loss $\mathcal{L}_{l1}$ and the GIOU (Rezatofighi et al., 2019) loss $\mathcal{L}_{giou}$ for bounding box regressions, and the cross-entropy classification loss $\mathcal{L}_{cls}$ for confidence scores as the joint optimization objective:

$$\mathcal{L}_{img} = \alpha\mathcal{L}_{l1} + \beta\mathcal{L}_{giou} + \gamma\mathcal{L}_{cls}, \tag{9}$$

where α, β and γ are the hyper-parameters to scale the losses, $\mathcal{L}_{img}$ is the visual target loss. Finally, the overall training loss $\mathcal{L}$ for this stage is:

$$\mathcal{L} = \mathcal{L}_{img} + \hat{\mathcal{L}}_{text}. \tag{10}$$

Model Training. We implement model training following a coarse-to-fine paradigm: 1) Pre-training on the coarser-grained MSD task by minimizing $\mathcal{L}_{text}$, and 2) Fine-tuning on the finer-grained MSTI task by minimizing $\mathcal{L}$, where the auxiliary task MSD is the predecessor training phase of the goal task MSTI. To this end, we unify the classification task for MSD and the sequence tagging task for textual target identification in MSTI into a text generation task. Note that for model testing on MSD, we use the model parameters obtained after the coarser-grained pre-training; in terms of the goal task MSTI, we directly input the test sarcastic sample into our finer-grained fine-tuned model to identify multimodal sarcasm targets.

	Acc.	P	R	F1
Att-BERT	80.03	76.28	77.82	77.04
CMGCN	79.83	75.82	78.01	76.90
HKE	76.50	73.48	71.07	72.25
DynRT	72.06	71.79	72.18	71.98
Multi-view CLIP	84.31	79.66	85.34	82.40
CofiPara-MSD	**85.70**	**85.96**	**85.55**	**85.89**

Table 1: Multimodal sarcasm detection results.

4 Experiments

4.1 Experimental Setup

Datasets. Our experiments are conducted based on two publicly available multimodal sarcasm datasets for evaluation: MMSD2.0 (Qin et al., 2023) and MSTI (Wang et al., 2022). Specifically, MMSD2.0 is a correction version of the raw MMSD dataset (Cai et al., 2019), by removing the spurious cues and fixing unreasonable annotation. In the coarser-grained pre-training stage, we utilized the large-scale MMSD2.0 dataset to pre-train our model for multimodal sarcasm detection. For the MSTI dataset, as the visual sarcasm target identification in the original MSTI data is mixed with optical character recognition, the reproduction of the code and data released by MSTI[2] has a big gap compared with the object detection results reported in their paper (Wang et al., 2022). Thus we introduce a refined version, i.e., MSTI2.0, to address the low-quality issue of the raw MSTI data by removing the visual target labels in images of only characters and converting them into the textual sarcasm target labels. Afterwards, MSTI2.0 is employed to fine-tune and evaluate the model in the finer-grained fine-tuning stage of our framework.

Baselines. We compare our model with the following multimodal baselines for multimodal sarcasm detection, which is the auxiliary task: 1) Att-BERT (Pan et al., 2020); 2) CMGCN (Liang et al., 2022); 3) HKE (Liu et al., 2022); 4) DynRT-Net (Tian et al., 2023); 5) Multi-view CLIP (Qin et al., 2023). We adopt Accuracy, F1 score, Precision, and Recall to evaluate the MSD performance.

To evaluate our model in multimodal sarcasm target identification that is our goal task, we compare the following state-of-the-art MSTI systems: 1) BERT-Base (Devlin et al., 2019); 2) BERT-Large; 3) Mask R-CNN (He et al., 2017); 4) YOLOv8 (Terven and Cordova-Esparza, 2023); 5) OWL-ViT (Minderer et al., 2022); 6) Grounding DINO (Liu et al., 2023b); 7) MSTI-RB (Wang et al., 2022); 8) MSTI-VB; 9) MSTI-CB; 10) MSTI-

[2] https://github.com/wjq-learning/MSTI

CL. We use Exact Match (EM) (Joshi et al., 2018) and F1 score (Molla and Joshi, 2019) as evaluation metrics of textual sarcasm target identification; and Average Precision (AP) (Lin et al., 2014), i.e., the COCO-style AP, AP50, and AP75, as the metrics for visual sarcasm target identification.

The data statistics, construction details of MSTI2.0, baseline descriptions and model implementation are detailed in the Appendix.

4.2 Main Results

Sarcasm Detection Performance. Table 1 illustrates the performance (%) of our proposed method versus all the compared representative multimodal baselines on the auxiliary task MSD. From these results, we have the following observations: 1) Compared to graph-based methods such as CMGCN and HKE and routing-based DynRT, Att-BERT that relies on semantic understanding has better performance, indicating that this task requires models to capture deep semantic information rather than superficial attributes. 2) Multi-view CLIP shows an overall advantage in its ability to align textual and visual features, and the isomorphic structures of text and image encoder also contribute to its superiority. 3) Our proposed CofiPara-MSD surpasses the leading baseline by 1.39% and 3.49% in accuracy and F1 score, additionally demonstrating a more balanced performance in terms of recall and precision, despite not primarily targeting the MSD task. The distinctive advantage of our model lies in the fact that while all the baselines solely focus on recognition, our model is equipped with rationales from divergent thinking with LMMs, which empowers our model to effectively uncover sarcastic content by adeptly leveraging the interplay between seemingly unrelated textual and visual elements within sarcasm.

Target Identification Performance. Table 2 shows the performance (%) of our method versus unimodal and multimodal baselines on the goal task MSTI. It can be observed that: 1) The unimodal methods, like text-modality models in the first group and image-modality models in the second group, fall short in simultaneously identifying both visual and textual sarcasm targets compared to the multimodal methods in the third group. 2) The textual target identification performance of visual grounding models (i.e., OWL-ViT and Grounding DINO), is hindered by the discrepancy between the MSTI task and their original pre-training objectives. Additionally, the lack of a consistent one-to-one correspondence between textual and visual targets in MSTI samples further contributes to their suboptimal performance. 3) Our method drastically excels in EM and AP50 compared to baselines, especially in visual target identification. We observe that CofiPara-MSTI improves textual target identification performance by 3.26% on average EM score compared to MSTI-VB, suggesting that our model is more precise in discerning sarcasm targets in the text modality of multimodal contents. On the other hand, our model exhibits a substantial superiority in visual target identification performance, especially on the AP50 metric, for an average improvement of 14.88% over the best visual performed baseline, indicating that our model can capture the correct visual targets within the image modality that contain sarcastic meanings, while baseline models perform poorly by simply identifying object rather than sarcasm targets, which further implies that our model displays a better understanding of multimodal sarcasm.

4.3 Ablation Study of Target Identification

As MSTI is our goal task, we conduct ablative studies on MSTI2.0 test data with the following variants: 1) *w/o MSD*: Simply train our model on the MSTI task without knowledge from pre-training on MSD. 2) *w/o LMM*: Use model parameters initialized by pre-training on the MSD task, and fine-tune directly on the MSTI task without knowledge from LMMs. 3) *w/o MSD&LMM*: Train our model directly on the MSTI task without any knowledge of LMM reasoning and MSD pre-training.

As demonstrated in Table 3, our model shows different degrees of performance degradation when MSD pre-training or LMM reasoning knowledge is ablated, indicating the effectiveness of our proposed method. Specifically, visual target identification performances show significant degradations by 2.10% and 13.13% on AP50 for *w/o MSD* and *w/o LMM* settings, respectively. This indicates that both LMM reasoning and MSD pre-training are helpful in identifying sarcasm targets, and that external LMM knowledge has a relatively larger impact on visual performance. We also notice that, the *w/o LMM* setting has relatively mild improvement over *w/o MSD&LMM*. This can be attributed to the fact that although the MSD pre-training itself may not necessarily significantly enhance model performance on the MSTI task with a large margin, it could help our model learn to implicitly ignore useless expressions and extract informative signals

	Dev					Test				
	EM	F1	AP	AP50	AP75	EM	F1	AP	AP50	AP75
BERT-Base	26.82	45.23	/	/	/	26.01	46.64	/	/	/
BERT-Large	29.29	46.42	/	/	/	27.89	46.93	/	/	/
Mask R-CNN	/	/	06.90	13.30	05.70	/	/	07.60	14.30	07.30
YOLOv8	/	/	06.58	12.81	06.13	/	/	10.49	17.57	11.18
OWL-ViT	14.80	01.20	03.36	13.75	00.17	18.40	01.64	03.32	14.47	00.91
Grounding DINO	18.29	01.60	11.15	19.77	10.37	15.22	00.59	10.92	17.26	11.30
MSTI-RB (ResNet+BERT-Base)	27.09	47.28	01.82	06.71	00.14	28.84	47.05	02.11	07.80	00.30
MSTI-VB (VGG19+BERT-Base)	28.19	45.74	02.03	07.43	00.27	29.51	49.02	02.57	08.92	00.24
MSTI-CB (CSPDarkNet53+BERT-Base)	27.62	48.00	03.78	13.68	00.40	27.89	48.39	03.80	13.06	01.03
MSTI-CL (CSPDarkNet53+BERT-Large)	28.18	48.32	02.64	09.56	00.86	28.70	49.78	02.80	11.02	00.91
CofiPara-MSTI	**31.96**	**49.53**	**15.38**	**34.29**	**15.57**	**32.26**	**50.27**	**13.79**	**32.49**	**12.01**

Table 2: Multimodal sarcasm target identification results.

	EM	F1	AP	AP50	AP75
CofiPara-MSTI	32.26	50.27	13.79	32.49	12.01
w/o MSD	30.24	49.61	13.72	30.39	12.15
w/o LMM	30.91	48.32	07.50	19.36	04.22
w/o MSD&LMM	30.10	50.72	06.34	17.21	04.61

Table 3: Ablation results on MSTI2.0 test set.

in the rationales from LMMs, highlighting its synergistic complementary with the LMM knowledge.

4.4 Case Study of Explainability

To better understand the mechanism of how LMM-generated rationales facilitate sarcasm target identification, we conduct a case study on the correctly predicted samples for better sarcasm explainability, as shown in Figure 3, where visual sarcasm targets are annotated by green rectangles and textual sarcasm targets are highlighted in red italics.

In these examples, we observe that: 1) rationales generated by LMM help promote the connections between two modalities. As shown in Figure 3(a), we notice that in the generated rationale, the image is depicted as a photo of Narendra Modi, which is then linked to the man who makes a refusing gesture in the picture. By introducing the connection between the word "narendramodi" and the man in the image, the target can be more easily recognized by our model; 2) on the other hand, rationales can complement background messages that are not given in the original texts and images, including both common sense and political knowledge. For example, in Figure 3(a), the rationale first recognizes the man as Prime Minister of India, and then offers a correction "for the first time ever" to the non-standard abbreviation of "for d 1st time ever", which is further explained as an expression of sarcastic tone towards Modi. Similarly, in Figure 3(b), LMM interprets "mlk" as Martin Luther King Jr. Day, the day to memorize dissenters who fought for civil rights, while the fact that police are arrest-

Text: #*narendramodi* den for a #selfie for d 1st time ever! #*narendramodi* #bjp #india #bjp4india #modi fest #digitalindia #skillindiaemoji_206

(a) LMM Rationale: The tweet is sarcastic because it makes fun of the Indian Prime Minister, Narendra Modi, by suggesting that he is denying a selfie for the first time ever. The image shows Modi with his palm facing the camera, which is a common gesture used to deny or refuse a request. The use of the phrase "for the first time ever" implies that Modi has never taken a selfie before, which is highly unlikely given his public profile. Overall, the tweet is a humorous take on Modi's public image and his love for selfies.

Text: big shout-out to the cops on mlk day for keeping us safe from dissenters . can 't have that . # thinblueline

(b) LMM Rationale: The tweet is sarcastic because it is praising the police for keeping people safe from dissenters on Martin Luther King Jr.'s birthday, which is ironic because King was a prominent dissenter who fought for civil rights and social justice. The tweet also includes a hashtag that supports the police, but King's legacy was built on opposing police brutality and advocating for nonviolent resistance. The tweet's message is opposite to King's beliefs and his legacy, making it a sarcastic commentary on the current political climate and the way people use social media to express their opinions.

Figure 3: Examples of correctly identified samples.

ing the dissenter in the image is in conflict with the context that expresses thanks to police, as well as the hashtag *#thinblueline*. The sarcasm target in the image is explained as the unjust political situation for people who fight for human rights, which is depicted in the image but outside the text. In this way, the rich but implicit correlations between the sarcasm text and image could be explained in visualized targets and readable snippets, which are also potentially valuable for aiding human checkers in verifying the sarcasm. We provide error analysis and explainability evaluation in the Appendix.

5 Conclusion and Future Work

In this paper, we proposed a novel coarse-to-fine paradigm to decode the implicit meanings hidden beneath the surface of texts and images in multimodal sarcasm for MSTI, by leveraging rich prior knowledge from LMM reasoning and MSD pre-training. We first inspired divergent thinking with

LMMs to derive competing rationales for coarser-grained pre-training of a small language model on MSD. Then we conducted finer-grained fine-tuning of the model on MSTI. Comprehensive experiments and analyses confirm the advantages of our framework. Future efforts aim to enhance our research by focusing on explicitly extracting useful information from generated rationales, to further relieve the inherent bias and variation in LMMs.

Limitations

There are multiple ways to further improve this work:

- Overall, this work's explainability primarily hinges on the sample's multi-modalities to justify its sarcasm through the visualized targets and readable rationales. Nevertheless, it does not delve into the more profound aspect of explainability concerning the internal workings of neural models. Future efforts will aim to enhance our research by focusing on improving the interpretability of the model's architecture.

- Generally, the distribution drift in datasets over time is a potential limitation for almost all data-driven tasks, especially for the multi-modal sarcasm on social media. However, one of the contributions of this work is proposing a novel paradigm to leverage commonsense reasoning knowledge in LMMs and pre-training insights from MSD for the MSTI task. The proposed framework is general enough, which should still work with newly released stronger LMMs or new sarcasm data appearing on social media. For example, in the future, we could publish a plug-and-play interface to incorporate a broader range of LMMs into our framework.

- Although sarcasm is defined much differently with hatefulness or offensiveness in previous literature (Xiong et al., 2019), in the future, we would try to incorporate more of the related social media and sentiment analysis datasets beyond our task to further broaden the boundaries of this research, such as multimodal target identification of hatefulness, offensiveness, sexism, or cyberbullying, etc.

- This work mainly focuses on the coarse-to-fine paradigm incorporating competing rationales from LMMs, to implicitly alleviate the negative impact posed by the potential noise in LMM-generated rationales. We would further update our paradigm to explicitly distill and display useful information from LMMs and avoid several common deficiencies of existing language models, including hallucination and limited generalization, as much as possible.

Ethics Statement

All data of MSTI2.0 come from the original MSTI dataset (Wang et al., 2022), which is an open-source dataset available for academic research. Our re-annotation and rectification process is designed to be carried out automatically. All the data in this work only include text and image modalities and do not contain any user information on social media.

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.

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*.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts,

Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

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.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *arXiv preprint arXiv:2309.12247*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. 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*, pages 294–297.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Aditya Joshi, Pranav Goel, Pushpak Bhattacharyya, and Mark Carman. 2018. Sarcasm target identification: Dataset and an introductory approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024a. Towards explainable harmful meme detection through multimodal debate between large language models. In *The ACM Web Conference 2024*, Singapore.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023a. 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*.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024b. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2543–2556.

Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on twitter with claim-guided hierarchical graph attention networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10035–10047.

Hongzhan Lin, Pengyao Yi, Jing Ma, Haiyun Jiang, Ziyang Luo, Shuming Shi, and Ruifang Liu. 2023b. Zero-shot rumor detection with propagation structure via prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5213–5221.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *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.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

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*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *CoRR*, abs/2306.08568.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer.

Diego Molla and Aditya Joshi. 2019. Overview of the 2019 alta shared task: Sarcasm target identification. In *Proceedings of the the 17th annual workshop of the Australasian language technology association*, pages 192–196.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2019. Detecting target of sarcasm using ensemble methods. In *Proceedings of the the 17th annual workshop of the Australasian language technology association*, pages 197–203.

Jasabanta Patro, Srijan Bansal, and Animesh Mukherjee. 2019. A deep-learning framework to detect sarcasm targets. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pages 6336–6342.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. 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*, pages 658–666.

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.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Juan Terven and Diana Cordova-Esparza. 2023. A comprehensive review of yolo: From yolov1 to yolov8 and beyond. *arXiv preprint arXiv:2304.00501*.

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.

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*.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. 2022. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8164–8175.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

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.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Zhiwei Yang. 2023a. Wsdms: Debunk fake news via weakly supervised detection of misinforming sentences with contextualized social wisdom. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1525–1538.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1).

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, Lynne Edwards, et al. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2(0):1–7.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

MMSD2.0	Train	Dev	Test
Sentences	19816	2410	2409
Positive	9572	1042	1037
Negative	10240	1368	1372

Table 4: Statistics of the MMSD2.0 dataset.

MSTI/MSTI2.0	Train	Dev	Test
Textual ST	2266/2830	360/502	367/517
Visual ST	1614/860	402/239	419/220
Total	3546	727	742

Table 5: Statistic comparison between MSTI and MSTI2.0.

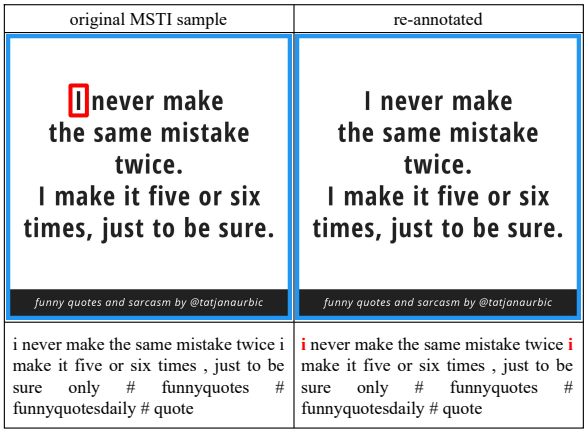


Figure 4: An example of re-annotated samples.

A Datasets

The statistics of MMSD2.0 and MSTI2.0 are shown in Table 4 and Table 5, respectively. In our experiments, we noted a significant discrepancy in the visual performance of the MSTI[3] baseline, trained using identical settings from their repository, compared to results reported in their original publication (Wang et al., 2022), as illustrated in Table 6. We dived into the quality of the MSTI dataset is poor due to the original MSTI dataset's substantial inclusion of visual labels targeting optical characters, as shown in the left column in Figure 4. We contend that labeling such entities as visual targets is not fitting for several reasons: 1) First, identifying optical character targets in synthetic images, such as memes or screenshots, fundamentally leans towards a Natural Language Processing (NLP) challenge. Applying visual metrics like AP50 to assess a task that is inherently textual in nature is clearly impractical. 2) Second, the recurrence of certain optical characters within a single image, such as the words "I", "you", or "Trump", often goes partially annotated in the original MSTI dataset, where typically only the first occurrence is marked, while subsequent instances are overlooked. This approach to annotation is also questionable. The statistic of optical character targets in the original MSTI dataset is listed in Table 7, where *OCR targets* denotes the optical character targets that consist of 48.53% of the visual labels.

We re-annotated the original dataset in the following steps: 1) First, by calculating the area ratios of truth-bounding boxes to images, we first roughly filter out samples that might contain OCR targets with a threshold of 0.15. 2) Then, by using the EasyOCR[4] tool, we extract texts in these images and their corresponding bounding boxes. 3) To further adjust the extracted results, we use LLaVA for result correction. Specifically, we feed the extracted texts and images into LLaVA, and prompt the model to check if the texts match the characters in the images. 4) Moreover, we applied a manual validation to double-check the corrected results, and finally concatenate the results to the original MSTI texts. The original bounding box labels are then deleted after being transformed into textual forms.

An example of the original and re-annotated sample is shown in Figure 4. The bounding box is labeled in the red rectangle, and textual labels are highlighted in red bold. The original sample on the left neglected the second "I" in the image, and the word "I" in the text also remains unlabeled, whereas in the re-annotated version, we removed the bounding box and labeled the target word in textual form.

B Baselines

We compare our model with the following baselines for multimodal sarcasm detection: 1) TextCNN (Kim, 2014): a neural network based on CNN for textual classification; 2) Bi-LSTM (Graves and Schmidhuber, 2005): a bi-directional long short-term memory network for textual classification; 3) SMSD (Xiong et al., 2019): a self-matching network to capture the incongruity information in sentences by exploring word-to-word interactions; 4) RoBERTa (Liu et al., 2019): an optimized version of the BERT (Devlin et al., 2019) language model; 5) ResNet (He et al.,

[3]https://github.com/wjq-learning/MSTI

[4]https://github.com/JaidedAI/EasyOCR

	Dev					Test				
	EM	F1	AP	AP50	AP75	EM	F1	AP	AP50	AP75
MSTI-CB (reported)	34.20	44.90	32.10	52.30	34.20	35.00	45.80	32.30	51.80	34.00
MSTI-CB (reproduced)	35.51	43.09	00.74	02.40	00.12	33.15	41.22	01.20	04.50	00.03
CofiPara-MSTI	39.37	45.90	04.68	09.79	03.95	40.04	44.43	04.22	09.62	03.54

Table 6: The gap between the reported MSTI-CB results in the original paper (Wang et al., 2022) and the reproduced MSTI-CB results by the released official codes on the original MSTI dataset. And we further provide the results of our proposed approach on the original MSTI dataset.

	Train	Dev	Test
OCR targets	754	163	199
non-OCR targets	860	239	220

Table 7: Statistics of optical labels in the original MSTI dataset.

2016): we employ the image embedding of the pooling layer of ResNet for sarcasm classification; 6) ViT (Dosovitskiy et al., 2020): a neural network model that applies the transformer architecture to computer vision tasks; 7) HFM (Cai et al., 2019): a hierarchical fusion model for multimodal sarcasm detection; 8) Att-BERT (Pan et al., 2020): two attention mechanisms is devised to model the text-tonly and cross-modal incongruity, respectively; 9) CMGCN (Liang et al., 2022): a fine-grained cross-modal graph architecture based on attribute-object pairs of image objects to capture sarcastic clues; 10) HKE (Liu et al., 2022): a hierarchical graph-based framework to model atomic-level and compositionlevel congruity; 11) DynRT-Net (Tian et al., 2023): a dynamic routing transformer network via adapting dynamic paths to hierarchical co-attention between image and text modalities; 12) Multi-view CLIP (Qin et al., 2023): a multimodal model based on CLIP (Radford et al., 2021) to capture sarcasm cues from different perspectives, including image view, text view, and image-text interactions view.

We adopt Accuracy, F1 score, Precision, and Recall to evaluate the MSD performance. Although we make a comparison with the MSD baselines for the illustration of our universal framework, please note that the MSD task is just an auxiliary task but not our goal task.

To evaluate our model in multimodal sarcasm target identification, we compare the following state-of-the-art MSTI systems: 1) BERT-Base (Devlin et al., 2019): the base version of a classical auto-encoding pre-trained language model is applied to the sequence tagging task; 2) BERT-Large: the large version of the BERT model; 3) Mask R-CNN (He et al., 2017): an extension of Faster R-CNN (Girshick, 2015) that adds a branch for predicting segmentation masks on each Region of Interest, enabling pixel-level object instance segmentation; 4) YOLOv8 (Terven and Cordova-Esparza, 2023): an advanced version of the YOLO (You Only Look Once) series, designed for real-time object detection with improved accuracy and speed; 5) OWL-ViT (Minderer et al., 2022): a pre-trained vision-language model designed for open-vocabulary object detection; 6) Grounding DINO (Liu et al., 2023b): a pre-trained vision-language model that could act as a powerful zero-shot object detector in the open set; 7) MSTI-RB (Wang et al., 2022): a variant of the MSTI model with ResNet152 as the backbone and BERT-Base as the language model; 8) MSTI-VB: a variant of the MSTI model with VGG19 as the backbone and BERT-Base as the language model; 9) MSTI-CB: a variant of the MSTI model with CSP-DarkNet53 as the backbone and BERT-Base as the language model; 10) MSTI-CL: a variant of the MSTI model with CSPDarkNet53 as the backbone and BERT-Large as the language model.

We use Exact Match (EM) accuracy (Joshi et al., 2018) and F1 score (Molla and Joshi, 2019) as evaluation metrics of textual sarcasm target identification; and Average Precision (AP) (Lin et al., 2014), i.e., the COCO-style AP, AP50, and AP75, as the metrics for visual sarcasm target identification.

Please note that due to our inability to evaluate all potential models, we can only select the representative state-of-the-art baselines for the text-modality, image-modality, and multimodal types.

C Implementation Details

In our experiments, we utilize two representative LMMs, LLaVA (Liu et al., 2023a) and Qwen-VL (Bai et al., 2023) to generate rationales, for their ex-

cellence in tasks like multimodal question answering and reasoning. Specifically, we implement the "LLaVA-1.5 13B" and "v1.0.0" versions for LLaVA and Qwen-VL, respectively. Note that the choice of LMMs is orthogonal to our proposed paradigm, which can be easily replaced by any existing LMM without further modification. Our CofiPara utilizes T5 architecture (Raffel et al., 2020) as our text encoder and decoder backbone. Specifically, we use the "flan-t5-base" version to initialize our language model. For image encoder, bi-attention block, and cross-modality image decoder structure, we utilize the same setting as GroundingDINO-T [5], and initialize using SwinT_OGC checkpoint in the same repository, where we freeze image encoder parameters during the whole training process. For text decoding, we adapt a text-to-image cross-attention module that matches text query with image key and value, which is randomly initialized. Hyper-parameters used during training are listed in Table 8, where these hyper-parameters are shared during both pre-training and fine-tuning process. Our results are averaged over 10 random runs. All of our experiments are conducted on a single NVIDIA RTX A6000 GPU. The size of our model is about 320M, among which 295M parameters are trainable. Training takes about 1 hour per epoch at the pre-training stage, and about 30 minutes per epoch at the fine-tuning stage. For MSTI baseline implementation, we use the exact same settings for training as reported in their paper (Wang et al., 2022). As for the OWL-ViT baseline, we use the specific version of "OWLv1 CLIP ViT-B/32", and train the model for 10 epochs using a batch size of 8, a learning rate of 5e-5. For the Grounding DINO baseline, we also implement the SwinT_OGC version, and we report the zero-shot results as the released version only supports zero-shot inference in their current repository without training codes.

D CofiPara Algorithm

Algorithm 1 presents the coarse-to-fine training paradigm of our approach.

First, we perform divergent thinking strategy on two datasets D_1 and D_2 for the pre-training and fine-tuning stages, respectively. For each MSD sample $M = \{I, T\}$ in D_1, we generate a rationale pair including a negative (i.e., `non-sarcastic`) and a positive (i.e., `sarcastic`) one with potential label $*$ that explicates the input as sarcastic/non-sarcastic. For MSTI samples in D_2, we only generate a single rationale r_{pos} that explains why the input sample is sarcastic. Then, we utilize the dataset D_1 for coarser-grained pre-training, where the model is trained to identify if the input sample M is sarcastic or not, and use the cross-entropy loss as the training loss. After $epoches$ of iteration for pre-training, we further fine-tune our model on the dataset D_2, to identify textual and visual sarcasm targets in the inputs. In the fine-tuning stage, the model outputs the predicted textual and visual target $\hat{y}_{text}, \hat{y}_{img}$, with which we calculate text and image loss as shown in Equation 7 and Equation 9. Finally, by adding up text loss $\hat{\mathcal{L}}_{text}$ and image loss $\mathcal{L}_{img}$, we optimize our model with the loss $\mathcal{L}$.

[5] https://github.com/IDEA-Research/GroundingDINO

Hyper-Parameters	
Epoch	10
Batch Size	4
Learning Rate	5e-5
Optimizer	Adam
Adam eps	1e-8
Image Size	600
L1 coefficient α	0.2
GIOU coefficient β	1e-3
Classification coefficient γ	0.1
Number of text decoder layers L	12
Number of image decoder layers K	6

Table 8: Hyper-parameters used during pre-training and fine-tuning.

	Acc.	P	R	F1
TextCNN	71.61	64.62	75.22	69.52
Bi-LSTM	72.48	68.02	68.08	68.05
SMSD	73.56	68.45	71.55	69.97
RoBERTa	79.66	76.74	75.70	76.21
ResNet	65.50	61.17	54.39	57.58
ViT	72.02	65.26	74.83	69.72
HFM	70.57	64.84	69.50	66.88
Att-BERT	80.03	76.28	77.82	77.04
CMGCN	79.83	75.82	78.01	76.90
HKE	76.50	73.48	71.07	72.25
DynRT	72.06	71.79	72.18	71.98
Multi-view CLIP	84.31	79.66	85.34	82.40
CofiPara-MSD	**85.70**	**85.96**	**85.55**	**85.89**

Table 9: Multimodal sarcasm detection results with text-modality, image-modality, and multimodal baselines.

E More Results of Sarcasm Detection Performance

Table 9 illustrates the performance (%) of our proposed method versus all the compared sarcasm detection baselines on the auxiliary task MSD ($p < 0.05$ under t-test). From the results, we

Algorithm 1 Model Training

Input: MSD dataset D_1, MSTI dataset D_2, template p, LMM, our model f

Rationale Generation:

for $M = \{I, T\} \in D_1, D_2$ **do**
  if $M \in D_1$ **then**
    $r_{pos}, r_{neg} = LMM(p(I, T, *))$
    $\hat{T} = \{T, r_{pos}, r_{neg}\}$
  end
  else
    $r_{pos} = LMM(p(I, T, *=\texttt{sarcastic}))$
    $\hat{T} = \{T, r_{pos}\}$
  end
end

Pre-training Stage:

while $epoch < epoches$ **do**
  for $M = \{I, \hat{T}, y\} \in D_1$ **do**
    $\hat{y} = f(I, \hat{T})$
    $\mathcal{L}_{text} = CE(\hat{y}, y)$
  end
end

Fine-tuning Stage:

while $epoch < epoches$ **do**
  for $M = \{I, \hat{T}, y\} \in D_2$ **do**
    $\hat{y}_{text}, \hat{y}_{img} = f(I, \hat{T})$
    calculate $\hat{\mathcal{L}}_{text}$ as Equation 7.
    calculate $\mathcal{L}_{img}$ as Equation 9.
    $\mathcal{L} = \mathcal{L}_{img} + \hat{\mathcal{L}}_{text}$
  end
end

	Acc.	P	R	F1
LLaVA (Zero-shot)	51.06	40.09	46.40	43.02
LLaVA (CoT)	48.69	40.93	65.17	50.28
LLaVA (Divergent-thinking)	80.41	82.22	82.55	80.40
LLaVA (CofiPara)	84.51	84.58	84.17	84.43
Qwen-VL (Zero-shot)	76.63	67.44	68.53	69.03
Qwen-VL (CoT)	68.86	66.82	68.67	66.96
Qwen-VL (Divergent-thinking)	75.88	73.68	77.19	74.12
Qwen-VL (CofiPara)	85.70	85.96	85.55	85.89
CofiPara-MSD w/o LMM	81.61	81.92	81.36	81.46

Table 10: Effect of different LMMs on MMSD2.0 dataset.

have the following observations: 1) The multimodal methods in the third group generally outperform the unimodal methods of text and image modality in the first and second groups, respectively. 2) Compared to graph-based methods such as CMGCN and HKE and routing-based DynRT, models like RoBERTa and Att-BERT that rely on semantic understanding have much better performance, indicating that this task requires models to capture deep semantic information rather than superficial attributes. 3) Multi-view CLIP shows an overall advantage in its ability to align textual and visual features, and the isomorphic structures of text and image encoder also contribute to its superiority. 4) Our proposed CofiPara-MSD, as a by-product of the coarse-to-fine framework, surpasses the leading baseline by 1.39% and 3.49% on the accuracy and F1 score, additionally demonstrating a more balanced performance in terms of recall and precision, despite not primarily targeting the MSD task. The distinctive advantage of our model lies in the fact that while all the baselines solely focus on recognition, our model is equipped with rationales from divergent thinking with LMMs, which empowers our model to effectively uncover sarcastic content by adeptly leveraging the interplay between seemingly unrelated textual and visual elements within sarcasm.

F LMM for MSD

Table 10 shows the effect of different LMMs (i.e., LLaVA (Liu et al., 2023a) and Qwen-VL (Bai et al., 2023)) on the auxiliary task MSD. The 'Divergent-Thinking' prompting strategy could effectively enhance the multimodal sarcasm detection performance of LMMs (Lin et al., 2024b), especially LLaVA, which suggests that the conflicting rationales generation from the divergent thinking is a reasonable way to optimize the reasoning chains

for LMMs applied to the MSD task. We notice that our model performance dropped on accuracy and F1 score with LMM-generated rationales removed. This suggests that LMM-generated rationales do provide supportive knowledge for detecting sarcasm, and that our model can effectively filter out useful information from contrary rationales. This design makes our model robust in the following fine-tuning stage for MSTI, which could alleviate the negative impact posed by potential noise in the rationale from only the sarcastic perspective.

G LMM for MSTI

Table 11 shows the effect of different LMMs (i.e., LLaVA (Liu et al., 2023a) and Qwen-VL (Bai et al., 2023)) on the goal task MSTI, to enhance the comprehensiveness and robustness of the evaluation with different prompting strategies: 1) Zero-shot: Directly prompt a representative LMM, to identify sarcasm targets; 2) CoT: Prompt the LMMs with the Chain-of-Thought reasoning; 3) Sarcastic: Prompt the LMMs with the generated rationales by itself from the sarcastic position; 4) CofiPara: Our proposed paradigm under full setting based on the integration of the reasoning knowledge from LMMs and the pre-training knowledge from MSD, where LMMs are LLaVA or Qwen-VL. Note that LLaVA only supports to perform textual sarcasm target identification with multimodal input. We can see from the evaluations of broader LMM-based models like LLaVA and Qwen-VL that, the direct deployment based on LMMs in the zero-shot, CoT or Sarcastic settings, struggles without lightweight design specific to the MSTI task, also implying the importance of divergent thinking with LMMs to alleviate the inherent bias and variation during deductive reasoning for multimodal sarcasm target identification.

H Effect of LMM-generated Rationale

As this work is the first to introduce LMMs into the multimodal sarcasm research area, all the MSD and MSTI baselines are not LMM-based. Although incidentally achieving such a more competitive performance on the MSD task as a by-product, our coarse-to-fine paradigm essentially focuses on the MSTI task. Please note that our proposed divergent thinking with LMM is orthogonal to the conventional MSD approaches. Although integration of our proposed LMMs' divergent thinking with the conventional MSD methods may lead to further performance improvements, it is not our research focus in this work. As all the MSD baselines cannot be applied directly to the goal task (i.e., multimodal sarcasm target identification), simply adding the LMM-generated rationales as input of the MSD baselines does not have any practical significance towards the goal task MSTI.

In contrast, we argue that it is necessary to augment the representative MSTI baselines with LMM-generated rationales for a fair comparison with our proposed framework. We first present Table 12 to illustrate how the two representative multimodal MSTI baseline models (i.e., Grounding DINO and MSTI-CB) would perform if equipped with LMM-generated rationales. One straightforward and possible solution would be: using the rationale generated from LMMs as input for an end-to-end training manner. Consequently, we use the rationale generated from the sarcastic perspective as the additional input for the baselines, similar to our paradigm in the finer-grained fine-tuning stage. We can see that both of the baselines cannot perform well after being equipped with the reasoning knowledge from LMMs. By this design, there may be two inherent weaknesses: 1) The knowledge elicited directly from LMMs may exhibit variation and bias; 2) If without our proposed divergent thinking mechanism with pre-training on MSD, the quality of generated rationales from LMMs needs some additional designs to filter. This reaffirms the advantage and necessity of our proposed coarse-to-fine paradigm for multimodal sarcasm target identification with LMMs.

Then we delve into the effect of rationales in our different training stages as shown in Table 13: 1) *CofiPara*-MSTI denotes using rationales generated from Qwen-VL at both training stages of coarser-grained pre-training and finer-grained fine-tuning, where pre-training uses the competing rationales and fine-tuning only uses the rationale from the sarcastic position; 2) *w/o RF* denotes using competing rationales at pre-training and no rationale at fine-tuning; 3) *w/o RP* denotes using the sarcastic rationale at fine-tuning without rationales as input at pre-training; 4) *w/o LMM* denotes using no LMM-generated rationale at both stages. It can be demonstrated that the variants after removing rationales suffer different degrees of performance degradation, indicating the effectiveness of our design philosophy for multimodal sarcasm target identification by divergent thinking with LMM and multimodal fusion with small LM.

	Dev					Test				
	EM	F1	AP	AP50	AP75	EM	F1	AP	AP50	AP75
LLaVA (Zero-shot)	16.05	00.14	/	/	/	13.20	00.40	/	/	/
LLaVA (CoT)	16.23	00.01	/	/	/	13.20	00.53	/	/	/
LLaVA (Sarcastic)	17.19	00.28	/	/	/	13.34	00.39	/	/	/
LLaVA (CofiPara)	32.92	50.47	12.26	32.21	10.11	33.46	50.91	12.84	31.29	10.88
Qwen-VL (Zero-shot)	18.98	00.49	00.30	01.32	00.08	14.82	01.35	00.32	01.43	00.06
Qwen-VL (CoT)	18.70	00.49	00.68	02.57	00.13	15.09	01.48	00.62	02.65	00.07
Qwen-VL (Sarcastic)	18.99	00.49	00.57	02.58	00.13	14.69	01.49	00.63	02.37	00.07
Qwen-VL (CofiPara)	31.96	49.53	15.38	34.29	15.57	32.26	50.27	13.79	32.49	12.01

Table 11: Effect of different LMMs on MSTI2.0 dataset.

	Dev					Test				
	EM	F1	AP	AP50	AP75	EM	F1	AP	AP50	AP75
Grounding DINO (LLaVA)	18.41	00.32	06.82	10.69	06.47	14.82	00.44	05.28	09.89	04.76
Grounding DINO (Qwen-VL)	18.15	00.01	05.20	08.93	04.18	14.55	00.44	05.23	10.69	04.87
MSTI-CB (LLaVA)	27.78	44.31	02.57	08.53	00.05	29.24	48.15	01.55	06.24	00.05
MSTI-CB (Qwen-VL)	28.61	44.81	00.89	04.49	00.02	29.11	47.32	00.86	05.05	00.01
CofiPara-MSTI (LLaVA)	32.92	50.47	12.26	32.21	10.11	33.46	50.91	12.84	31.29	10.88
CofiPara-MSTI (Qwen-VL)	31.96	49.53	15.38	34.29	13.39	32.26	50.27	13.79	32.49	12.01

Table 12: The impact of rationales generated by different LMMs on different MSTI baselines.

	Dev					Test				
	EM	F1	AP	AP50	AP75	EM	F1	AP	AP50	AP75
CofiPara-MSTI	31.96	49.53	15.38	34.29	13.39	32.26	50.27	13.79	32.49	12.01
w/o RF	28.80	45.75	09.10	22.15	06.18	30.24	49.59	08.82	19.75	07.25
w/o RP	28.80	49.54	11.09	25.98	08.27	31.72	50.69	10.14	21.86	09.15
w/o LMM	31.13	47.30	08.87	21.09	06.76	30.91	48.32	07.50	19.36	04.22

Table 13: The influence of rationales at different training stages.

I Detailed Analysis of Competing Rationales

To further analyze how the competing rationales applied in the pre-training stage affect the model training and decision-making process, we conduct more detailed and comprehensive ablation studies[6] on the competing rationales derived from Qwen-VL due to its best-performed results, as shown in Table 14 and Table 15. In the results of coarser-grained pre-training in Table 14, *w/o* r_{neg} refers to only using the rationale r_{pos} that explains why the sample is sarcastic as part of the input in the pre-training stage, likewise, *w/o* r_{pos} indicates only using the rationale r_{neg} that explains why the sample is non-sarcastic as part of the input in the pre-training stage. For finer-grained fine-tuning results in Table 15, due to the nature of the MSTI task as emphasized in §3.3, only the same rationale r_{pos} from the sarcastic position can be used as part of the input in the fine-tuning stage, where *w/o* r_{neg} and *w/o* r_{pos} denote that the model parameters for fine-tuning are initialized from the models pre-trained on corresponding settings in Table 14, respectively.

From Table 14, we observe that compared to CofiPara-MSD that uses competing rationales, both *w/o* r_{neg} and *w/o* r_{pos} showed sharp decreases in sarcasm detection, suggesting that introducing either r_{pos} or r_{neg} alone could bring severe noises that are one-way grossly misleading. The negative impact of adding r_{neg} is larger than r_{pos}, as r_{pos} relatively provides more information, which is instinctively understandable. In the fine-tuning stage, as shown in Table 15, by removing different rationales at the pre-training stage, the models showed performance drops to varying extents. The fact that *CofiPara*-MSTI has the best performance compared to other settings proves that competing rationales in pre-training indeed make our model more robust to noisy LMM rationales, as well as facilitating reasoning. On the other hand, *w/o* r_{pos} and *w/o* r_{neg} settings both outperform *w/o RP* that uses no rationales in pre-training, showing that using either rationale could bring a certain level of improvement in robustness, which accords with our naive idea. Moreover, the competing rationales could further encourage divergent thinking for multimodal sarcasm by resorting to a fundamental characteristic of human problem-solving, i.e., competing statements. Overall, the comprehensive results reaffirm that our proposed competing rationales for coarser-grained pre-training could better contribute to multimodal sarcasm target identification.

[6]Note that as the relative positions between the competing rationales are set as invariant, our preliminary verification indicated that the order information has minimal impact on the model's learning process. Therefore, here we do not delve further into this aspect.

	Acc.	P	R	F1
CofiPara-MSD	85.70	85.96	85.55	85.89
w/o LMM	81.61	81.92	81.36	81.46
w/o r_{neg}	68.55	68.68	68.49	68.99
w/o r_{pos}	58.57	51.94	53.28	49.10

Table 14: The influence of competing rationales on multimodal sarcasm detection.

J Effect of Different Proposed Components

Besides the core ablative test results on the MSTI2.0 test set shown in Table 3 and the supplemental results on the dev set shown in Table 17, we further perform ablation studies by discarding some important components of our model: 1) *w/o QSM*: Use global visual features instead of query-selected local features for cross-modality text decoding. 2) *w/o CTD*: Remove the image-to-text cross-attention in the text decoder, and only use textual features for text decoding. 3) *w/o CID*: Remove the image-to-text and text-to-image cross-attention in the image decoder, and only use self-attention in image decoding.

As demonstrated in Table 16, it can observed that, without using the query-selection mechanism, our model showed a slight improvement in EM and F1 score, while AP50 downgraded by 2.09% and 1.20%. This can be explained as, by using query-selected local features, our model indeed missed information in identifying text targets. However, the reason why query selection benefits visual performance, we speculate, is that by fusing local features that are more text-related, the text encoder of our model can learn representations that correlate more to the key components in the images, thus benefiting the follow-up image decoding process. The performance degradation in *w/o CTD* can be elucidated by similar reasons, the existence of cross-modality text decoder forced text encoder to learn representations that benefit cross-modality fusion in images, further facilitating image decoding. For *w/o CID* setting, we observed that with text-to-image and image-to-text cross-attention removed, there's a drastic drop in visual target identification

	Dev					Test				
	EM	F1	AP	AP50	AP75	EM	F1	AP	AP50	AP75
CofiPara-MSTI	31.96	49.53	15.38	34.29	13.39	32.26	50.27	13.79	32.49	12.01
w/o RP	28.80	49.54	11.09	25.98	08.27	31.72	50.69	10.14	21.86	09.15
w/o r_{neg}	31.82	49.19	12.57	26.23	12.32	31.85	49.41	12.63	28.15	10.94
w/o r_{pos}	31.13	49.54	14.99	31.85	11.72	29.97	50.39	13.11	26.45	10.43

Table 15: The influence of competing rationales on multimodal sarcasm target identification. *w/o* r_{neg} and *w/o* r_{pos} mean that the model parameters for fine-tuning are initialized from the pre-training stage under different settings, where only the rationale from sarcastic or non-sarcastic perspectives is used as part of the input, respectively, instead of our proposed competing rationales with both sarcastic and non-sarcastic positions.

performance, indicating that the model loses its object detection ability, which proves that cross-modality fusion in image decoding is essential.

K Evaluation of Rationale Quality

Automatic Evaluation. Generally, there is no gold explanation about multimodal sarcasm for the multimodal sarcasm target identification task due to the diverse forms of textual expression. Devising reliable metrics without reference is not a straightforward task and can also be problematic. Furthermore, different types of text necessitate the evaluation of distinct aspects, such as informativeness, fluency, soundness, etc. (Fabbri et al., 2021; Mehri and Eskenazi, 2020), which makes it hard to design metrics for each type of text and dimension separately. Nowadays, GPT-4V (OpenAI, 2023) has revolutionized the field of LMMs with a more powerful expressive capacity for multimodal inputs. In this subsection, we present a new automatic evaluation using GPT-4V in a reference-free mode, to evaluate the text quality of the explanations generated by our approach from LLaVA and Qwen-VL.

We randomly selected 50 samples from the MSTI2.0 test set. Specifically, GPT-4V is prompted to score the explanations w.r.t. each multimodal sample according to the following criteria: 1) *Conciseness*: the explanation contains less redundant information; 2) *Informativeness*: the explanation provides new information, such as explaining the background and additional context; 3) *Persuasiveness*: the explanation seems convincing; 4) *Readability*: the explanation follows proper grammar and structural rules; 5) *Soundness*: the explanation seems valid and logical. For each criterion, a 3-point Likert scale was employed, where 1 meant the poorest quality and 3 the best.

Table 18 demonstrates the averaged scores of the explanation evaluation by GPT-4V on the two sources (i.e., LLaVA and Qwen-VL) regarding the five criteria. We could observe that: 1) Qwen-VL scores higher than LLaVA on Conciseness, Informativeness and Persuasiveness scores, indicating that Qwen-VL is more effective in providing information, as well as providing convincing explanations. 2) LLaVA, however, excels in Readability and Soundness scores, which means that it has a better capability of reasoning. 3) We also notice that rationales generated by Qwen-VL are longer than those of LLaVA, with average lengths of 120.98 and 99.68, which further proves that Qwen-VL offers more evidence in its explanations.

Human Evaluation. Considering that automatic evaluation cannot realistically measure the quality of the chosen explanations generated by LMMs, we further conduct the human subjects study to evaluate the overall quality of explainability. For the 50 samples randomly selected in the previous automatic evaluation phase, 10 professional linguistic annotators (five females and five males between the ages of 26 and 29) are asked to evaluate the explanations of our model from LLaVA and Qwen-VL. The metrics of human evaluation are the same as the *automatic evaluation*.

To protect our human evaluators, we establish three guidelines: 1) ensuring their acknowledgment of viewing potentially uncomfortable content, 2) limiting weekly evaluations and encouraging a lighter daily workload, and 3) advising them to stop if they feel overwhelmed. Finally, we regularly check in with evaluators to ensure their well-being.

The scores of human evaluation are shown in Table 18. Note that the intra-class agreement score is 0.651. The average Spearman's correlation coefficient between any two annotators is 0.632. We can observe that: 1) Qwen-VL explanations are more human-preferable, beating LLaVA on almost every metric. 2) Qwen-VL mainly shows advantages

	Dev					Test				
	EM	F1	AP	AP50	AP75	EM	F1	AP	AP50	AP75
CofiPara-MSTI	31.96	49.53	15.38	34.29	13.39	32.26	50.27	13.79	32.49	12.01
w/o QSM	32.92	50.47	12.26	32.21	10.11	33.46	50.91	12.84	31.29	10.88
w/o CTD	30.58	49.35	12.91	26.48	11.31	32.25	51.87	11.96	25.44	11.34
w/o CID	31.82	49.87	00.18	00.83	00.02	31.72	51.84	00.38	01.76	00.04

Table 16: The influence of query-selection mechanism (*w/o QSM*), cross-modality text decoder (*w/o CTD*), and cross-modality image decoder (*w/o CID*).

	EM	F1	AP	AP50	AP75
CofiPara-MSTI	31.96	49.53	15.38	34.29	15.57
w/o MSD	31.41	48.78	13.15	28.28	12.14
w/o LMM	31.13	47.30	08.87	21.09	06.76
w/o MSD&LMM	29.21	48.51	08.82	23.92	05.07

Table 17: Ablation results on MSTI2.0 dev set. Results of the test set are provided in Table 3.

	GPT-4V		Human	
	LLaVA	Qwen-VL	LLaVA	Qwen-VL
Conciseness	2.00	2.52	2.29	2.31
Informativeness	2.08	2.38	2.45	2.60
Persuasiveness	2.08	2.34	2.02	2.22
Readability	2.77	2.54	2.69	2.69
Soundness	2.27	2.23	2.21	2.35

Table 18: Automatic GPT-4V and human evaluation of LMM-generated rationales on MSTI2.0 dataset.

in Informativeness, Persuasiveness and Soundness scores, denoting that its superiority lies in both reasoning and resolving sarcastic shreds of evidence. 3) According to our evaluators, Qwen-VL and LLaVA both show excellent fluency in the generated contents, where LLaVA tends to use clauses to make overall descriptions, while Qwen-VL favors short sentences that depict detailed facts. 4) We notice that human and automatic evaluations show deviation in Readability and Soundness. We conjecture GPT-4V prefers LLaVA over Qwen-VL for the reason that, LLaVA explicitly used GPT-4 generated instruction tuning data, which makes LLaVA rationales slightly more fluent and logical to the taste of GPT-4V. On the other hand, Qwen-VL applied a cleaning strategy on classical pre-training tasks, removing a large proportion of data. In addition, it added tasks like OCR and grounding, which may change its overall reasoning paths into patterns not so favorable to GPT-4V, but make sense from human perspectives.

L Error Analysis

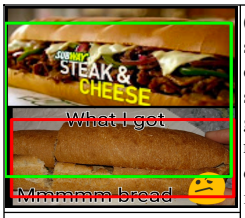
Apart from the case study in §4.4, we also conduct an error analysis for a better understanding of our proposed method. Figure 5 shows two of the incorrectly predicted examples[7], where the truth boxes in the image are labeled in red rectangles and the predicted boxes are the green ones. For text targets, we labeled our outputs in red italics, and ground truth in blue.

In the first example, the sarcasm in the original message refers to the poor sandwich that almost looks like two pieces of bread. While in LMM rationale, this tweet is over-analyzed as sarcastic towards SUBWAY's false advertisement, which misleads our model to generate the box that covers both the ad and the sandwich. Although LMM makes a good explanation of this example, the information it provides is partly unnecessary and may lead to false predictions.

As for the second example, where our model identifies a correct visual target but a wrong textual target. In the explanation, LMM misinterpreted the phrase "golf" as a word for a provocative image. Although the explanation reached the correct answer in the end that the sarcasm lies on the contrary between the "desiring" situation and the use of the word "hate", this situation is wrongly correlated to the word "golf", misleading our model to give a false prediction.

From the error analysis, we can also find that multimodal sarcasm target identification is more challenging than unimodal sarcasm target identification because it requires the integration and interpretation of heterogeneous data sources, i.e., texts and images, to understand the nuanced and often contradictory nature of sarcasm. This complexity demands sophisticated models capable of capturing and correlating features across different modalities

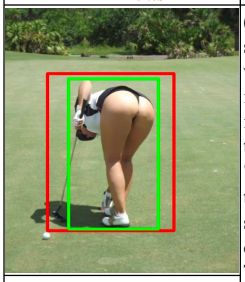
[7] **Disclaimer:** This section contains content that may be disturbing to some readers.



Text ground truth: <user> i can 't believe how close to the add my sandwich looks . what the hell is *this* ?

Our model: <user> i can 't believe how close to the add my sandwich looks . what the hell is *this* ?

(a) LMM Rationale: The tweet is sarcastic because the image shows a close-up of bread from a Subway sandwich, while the text says "What I got: mmmmm bread." The tweet is making fun of the fact that the image does not show any meat or other ingredients from the advertised Steak and Cheese sandwich. The tweet implies that the sandwich is only made of bread, which is a clear exaggeration. The use of sarcasm and humor in the tweet suggests that the person who created it is not satisfied with the actual contents of the sandwich and is poking fun at the misleading advertisement.



Text ground truth: don 't u just hate when ur at <user> and this happens to u ? ? ? # golf # sundayfunday emoji_3203 emoji_3205 emoji_92

Our model: don't u just hate when ur at <user> and this happens to u ? ? ? # *golf* # sundayfunday emoji_3203 emoji_3205 emoji_92

(b) LMM Rationale: The tweet is sarcastic because the image shows a woman bending over on a golf course in a revealing outfit, which is likely intended to be provocative. However, the text accompanying the image states "don't u just hate when ur at golf and this happens to u ? ? ? # golf # sundayfunday emoji_3203 emoji_3205 emoji_92", which is a play on words. The phrase *when ur at golf* is a double entendre, as it can be interpreted both as being on a golf course and being in a situation where a provocative image is displayed. The use of the word "hate" is also ironic, as the image is intended to be seen as desirable, rather than something to be disliked. Overall, the tweet is a form of sarcasm, using the image and text to make a commentary on the objectification of women in popular culture.

Figure 5: Examples of wrongly identified samples by our proposed CofiPara framework for MSTI.

to accurately identify the target of sarcasm, which is inherently more difficult than analyzing a single type of data.

M More Examples of Explainability

We present more examples in Figure 6 and Figure 7 to illustrate the supportive basis for the explanation of multimodal sarcasm. From the explanations, it can be observed that the LMM usually explains sarcasm in a certain mode, by describing first and then reasoning. In its descriptions, LMM can extract key information like political figures, historical scenes and trends in graphics, and by combing texts and images, it can also decipher some of the desensitized words like *<user>* that refer to a certain person in the context. We also noticed that, in its reasoning procedure, the LMM tends to perceive sarcasm by using the incongruity between texts and images, and the non-sarcastic samples are defined as non-contradictory, which is quite similar to human ways of thinking.

N Discussion about LMMs

In this section, we discuss potential concerns on LMMs in the following four aspects: 1) **Reproducibility**: Since our proposed method uses the open-source LMMs like LLaVA and Qwen-VL as our backbone model instead of the close-source GPT-4V, the results are definitely reproducible with open-source codes. In order to make sure the LLaVA or Qwen-VL can generate the same contents for the same instance, we utilized the specific version 'LLaVA-1.5 13B' of LLaVA and 'v1.0.0' of Qwen-VL, and further set the parameter temperature as 0 without sampling mechanism, that is, the greedy decoding was adopted to ensure the deterministic results for the explanation generation with the same prompt. 2) **High-quality Explanations**: Although both LLaVA and Qwen-VL are well-known excellent LMMs, there still could be noise and the quality of reasoning from them may vary. This is indeed a tricky issue for us to consider in this work, for which we have proposed both the 'Divergent Thinking with LMM' mechanism and the coarse-to-fine training paradigm to ensure the quality of reasoning utilized in our proposed model. The coarser-gained pre-training stage combined with the competing rationales derived from the divergent thinking of LMMs could facilitate our model to be more resilient and implicitly distill useful information in the following finer-grained fine-tuning stage. Furthermore, as illustrated in §K, we also make a systematic evaluation to assess the quality of the rationale explainability from both automatic and human perspectives. 3) **Test Set Leakage**: The test set leakage issue does not exist in LLaVA or Qwen-VL as their papers have clearly described the instruction-tuning data used for training, which does not include any multimodal sarcasm dataset used in our task. As shown in Table 10 and Table 11, where the LLaVA or Qwen-VL was directly deployed to test on the two standard datasets, LLaVA or Qwen-VL did struggle on the datasets. When replacing LLaVA with Qwen-VL in our framework, we can consistently observe enhanced performance. This suggests that such improvement is basically attributed to our designed framework rather than test set leakage, since directly prompting LLaVA or Qwen-VL should result in fairly good performance if it takes advantage of test set leakage. 4) **Generalizabil-**

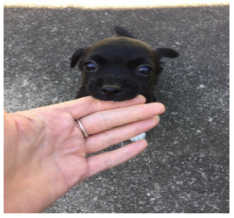
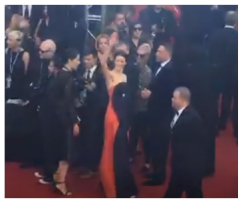
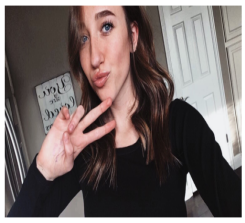
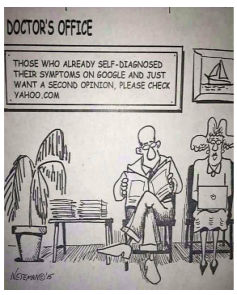
 **label**: non-sarcastic **text**: you taste like life-long friendship	**(A)** **pos**: The tweet is sarcastic because the text "you taste like life-long friendship" is accompanied by a photo of a small black puppy being held in someone's hand. The size of the puppy and the phrase "you taste like life-long friendship" suggest a romantic or emotional connection, but the image of a puppy being held in someone's hand is often used to convey a sense of cuteness and innocence. The contrast between the two ideas creates a humorous and ironic tone, indicating that the tweet is intended to be sarcastic. **neg**: The tweet is not sarcastic because the text "you taste like life-long friendship" and the image of a small black puppy being held in someone's hand create a sense of innocence and cuteness. The puppy's big eyes and small size add to the emotional appeal of the image, making it more likely to be interpreted as sincere rather than sarcastic. Additionally, the use of a puppy as the subject of the tweet suggests a desire to share a positive and uplifting message with the audience, further supporting the idea that the tweet is not intended to be sarcastic.
 **label**: non-sarcastic **text**: caitriona is here and she 's perfect omg	**(B)** **pos**: The tweet is sarcastic because the image shows a woman in a red dress on a red carpet, and the text accompanying the image is a sarcastic comment about the actress's appearance. The comment implies that the woman is not beautiful, which is a contradiction to the image of her on a red carpet surrounded by photographers and fans. The use of sarcasm suggests that the person who posted the tweet is mocking the idea that the woman is truly happy and enjoying her time at the event. **neg**: The tweet is not sarcastic because the text and the image match each other. The text says "caitriona is here and she 's perfect omg # goldenglobes # cheerstocait" and the image shows a woman in an orange dress on the red carpet. There is no indication of sarcasm in the content of the tweet.
 **label**: non-sarcastic **text**: i did something yesterday !	**(C)** **pos**: The tweet is likely sarcastic because the text "i did something yesterday!" is accompanied by a playful hand gesture, specifically the peace sign, and the user's selfie. The use of emojis and the playful tone of the text suggest that the user is being sarcastic or making a humorous statement. Additionally, the phrase "i did something yesterday" is a common phrase used to express a mundane or uninteresting activity, which further supports the idea that the user is being sarcastic. **neg**: The tweet is not sarcastic because the text "i did something yesterday!" and the image of a woman making a peace sign with her fingers are not contradictory or intended to be humorous in a sardonic manner. The text seems to be a straightforward statement, and the image is a common way to express peace or victory. There is no evidence in the image or text that suggests sarcasm or irony.
 **label**: sarcastic **text**: a poor indian politician busy in entertaining britishers !	**(D)** **pos**: The tweet is sarcastic because it depicts an image of an Indian politician laughing and joking with a British officer, despite the fact that India gained its independence from British rule in 1947. The caption "a poor Indian politician busy in entertaining Britishers" further emphasizes the irony and sarcasm of the situation. The tweet is making fun of the idea that Indian politicians are still subservient to the British and are not able to move on from their colonial past. **neg**: The tweet is not sarcastic because it is a historical photograph of Jawaharlal Nehru, the first Prime Minister of India, standing with Lord Louis Mountbatten, the last Viceroy of India, and Edwina Mountbatten, the wife of Lord Louis Mountbatten, in 1947. The tweet is sharing this photograph to celebrate the birth anniversary of Jawaharlal Nehru in Odia.
 **label**: sarcastic **text**: on one level up	**(E)** **pos**: The tweet is sarcastic because it makes fun of people who rely too heavily on online searches for medical advice. The cartoon shows a doctor's office with a patient who has already self-diagnosed their symptoms based on a Google search. The doctor is reading a book, implying that the patient is seeking a second opinion, but the patient's response is to check their symptoms on Yahoo.com. This is a play on the idea that people often rely too heavily on online sources for medical advice, rather than seeking a professional diagnosis from a doctor. The cartoon implies that this approach is not reliable and that seeking a second opinion from a qualified medical professional is necessary. **neg**: The tweet is not sarcastic because it is a cartoon that is making a point about people who rely on online searches for medical advice. The cartoon shows a doctor's office with a patient standing in front of the doctor, who is reading a book. The doctor is saying "Those who already self-diagnosed their symptoms on Google and just want a second opinion, please check Yahoo.com." The cartoon is not making a sarcastic comment, but rather making a commentary on the reliance on online search for medical advice.

Figure 6: Examples of competing rationales at the coarser-grained pre-training stage of our proposed CofiPara.

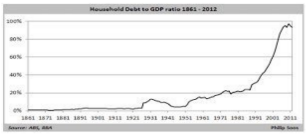
 **text**: <user> way to go timmies ! thanks for copious amount of cream cheese evenly spread	**(A)** **LMM**: The tweet is sarcastic because the image shows two slices of bagel with cream cheese, but the bagel appears to be broken in half. The broken bagel is placed on a napkin with the brand name Manhattan Bagel written on it. The fact that the bagel is broken and the cream cheese is spread unevenly implies that the person who took the photo did not have a good experience with the bagel. The person might be sarcastically thanking Manhattan Bagel for the poor quality of the bagel. The broken bagel could also be a symbol of the disappointment and frustration that the person feels towards the brand. The tweet is an example of a sarcastic expression where the person is expressing dissatisfaction with the product while using a positive phrase.
 **text**: i love it when they pay attention to the small details ! # breakingnews # supergirl	**(B)** **LMM**: The tweet is sarcastic because the image shows a news anchor delivering breaking news, but the text accompanying the image reads "I love it when they pay attention to the small details!" This is an ironic statement, as breaking news is typically associated with important and urgent events, and the phrase "small details" implies that the news anchor is more focused on minor details rather than the main story. The use of sarcasm in this tweet suggests that the person who created it is mocking the seriousness of news anchors and the perception that they are always delivering important information.
 **text**: i ' m just so relieved we have a <user> that will deal with north korea in a level-headed manner !	**(C)** **LMM**:The tweet is sarcastic because the image of an explosion with the caption "I'm just so relieved we have a [insert name of president] that will deal with North Korea in a level-headed manner!" is juxtaposed with an angry and aggressive president Trump pointing his finger. This creates a contrast between the positive statement about the president's ability to handle North Korea and the negative connotation associated with the angry and aggressive image of Trump. The use of sarcasm implies that the author is not serious about their statement and is instead making a humorous and critical commentary on the president's behavior and the situation with North Korea.
 **text**: oh joy ... first day of *summer school camp* and this just looks like so much fun !	**(D)** **LMM**: The tweet is sarcastic because the image shows a map of the Midwest with a large green area indicating heavy rain, but the text says "oh joy, first day of summer school/camp and this just looks like so much fun!" This implies that the person is not excited about summer school or camp, but is instead making a sarcastic comment about how much they enjoy these activities. The use of the phrase "oh joy" and the juxtaposition of the negative weather conditions with the positive sentiment in the text creates a sense of irony and sarcasm.
 **text**: <user> the *financialisation of our economy* is great for everyone ! ! ! haha <user> <user>	**(E)** **LMM:** The tweet is sarcastic because the statement "the financialisation of our economy is great for everyone!" is contradicted by the image and the text. The image shows a graph of the household debt to GDP ratio from 1861 to 2012, which indicates a steady increase in debt over time. This suggests that the financialization of the economy has not benefited everyone, but has instead led to an increase in debt. The sarcastic tone is implied by the use of the word "haha" at the end of the statement, which implies that the author is mocking the idea that financialization is beneficial to everyone.

Figure 7: Examples of sarcasm explainability of our proposed CofiPara.

ity: We believe our CofiPara framework is a general technique that works with LMMs, because our method does not choose GPT-4V as the representative LMM to generate rationales, but works well with the open-source LLaVA or Qwen-VL, which is not an OpenAI system.

O Dicussion about MSD and MSTI

Generally, for multimodal sarcasm moderation, MSD is the initial step of determining the presence of sarcasm in an expression at the holistic semantics (Qin et al., 2023), and MSTI goes a step further to analyze the specific focus of the sarcasm (Joshi et al., 2018). As MSD should be conducted first to establish the presence of sarcasm at the coarse level, followed by MSTI to explicitly identify the specific targets of the detected sarcasm at the fine level (Joshi et al., 2018), we devise a cohesive framework to operate on the coarse-to-fine training paradigm, aimed at pinpointing nuanced visual and textual targets of multimodal sarcasm, where multi-task learning is obviously not suitable due to their cascaded correlation.

P Future Work

We will explore the following directions in the future:

- Explainability evaluation: Although this study is at the forefront of using GPT-4V for the automated evaluation of rationale quality, the results exhibit slight discrepancies when compared to human subject studies. Furthermore, integrating GPT-4V into the rationale generation phase necessitates the exploration of even more advanced language models for evaluating the explanations produced by GPT-4V. Therefore, there is a need for more precise automatic evaluation methods for explanation quality. Simultaneously, conducting more extensive human subject studies with a broader pool of evaluators in a systematic manner would further validate our findings.

- LMM prompting: Our initial approach involved heuristically designing a single-turn prompt for LMMs to facilitate divergent thinking. However, we observed instances where the generated text overlooked key details, such as the semantic context of the sample. Moving forward, we plan to refine our prompting strategy to incorporate multi-turn or structured interactions (Lin et al., 2023b) with LMMs. This will enable more effective activation of commonsense reasoning related to vulnerable targets within sarcastic content, enhance visual feature extraction, and foster improved multimodal reasoning.

- High-quality benchmark: We plan to expand our research by developing additional high-quality MSTI benchmarks from perspectives like dataset construction on social media with propagation structure (Lin et al., 2021, 2022; Yang et al., 2023a), and updating our framework to incorporate a wider range of more powerful open-source LMMs, as they become available. This will enable us to further explore and evaluate the efficacy of LMMs in enhancing sarcasm awareness and understanding.