DeepMSD: Advancing Multimodal Sarcasm Detection Through Knowledge-Augmented Graph Reasoning

Yiwei Wei, Hengyang Zhou, Shaozu Yuan, Meng Chen, Haitao Shi, Zhiyang Jia, Longbiao Wang, *Member, IEEE*, and Xiaodong He, *Fellow, IEEE*

***Abstract*— Multimodal sarcasm detection (MSD) requires predicting the sarcastic sentiment by understanding diverse modalities of data (e.g., text, image). Beyond the surface-level information conveyed in the post data, understanding the underlying deep-level knowledge–such as the background and intent behind the data–is crucial for understanding the sarcastic sentiment. However, previous works have often overlooked this aspect, limiting their potential to achieve superior performance. To tackle this challenge, we propose DeepMSD, a novel framework that generates supplemental deep-level knowledge to enhance the understanding of sarcastic content. Specifically, we first devise a Deep-level Knowledge Extraction Module that leverages large vision-language models to generate deep-level information behind the text-image pairs. Additionally, we devise a Cross-knowledge Graph Reasoning Module to model how humans use prior knowledge to identify sarcastic cues in multimodal posts. This module constructs cross-knowledge graphs that connect deep-level knowledge with surface-level knowledge. As such, it enables a more profound exploration of the cues underlying sarcasm. Experiments on the public MSD dataset demonstrate that our approach significantly surpasses previous state-of-the-art methods.**

***Index Terms*— Multi-modal sarcasm detection, deep-level knowledge, large vision-language models, cross-knowledge graph.**

I. Introduction

THE sarcasm detection task aims to identify the sarcastic sentiment humans convey through their tweets.

In past years, the relevant research primarily focused on textual modalities [1], [2], [3], which explored the sarcasm contexts as additional cues to model the congruence level of texts, resulting in consistent improvement. With the development of the Internet, individuals are more willing to express their sentiments via multimodal data [4], [5], [6], including images and text, instead of purely text. Therefore, relying solely on text modality often fails to fully capture the emotions behind the post.

Compared to sarcasm detection in pure text, multimodal sarcasm detection is more effective and accurate, as it provides supplementary visual cues for sarcasm detection. Recent studies [7], [8], [9], [10], [11], [12] proposed various frameworks to integrate different modalities. To explore sarcastic cues within multimodal content, researchers have employed a diverse array of approaches, such as decomposition and relation networks [7], attention mechanisms [13], [14], graph-based methods [9], [11], [15], [16] and incongruity perceiving network [12]. These works effectively uncover the contradiction between the image and text modalities and further improve the performance of MSD tasks.

Nevertheless, previous studies have primarily focused on the surface-level knowledge conveyed in posts, such as the images and text within tweets. These elements alone are insufficient to capture the underlying emotional information embedded in the tweets. When humans understand the ironic emotions of the posts, it is necessary to associate deep-level knowledge, such as the background and intent behind the post. As the example illustrated in Figure 1, the post indicates deep-level knowledge, stating that "*the tweet appears to be humorous . . . commentary on the political situation in the United States.*". This knowledge encompasses a broad range of world facts [17] about public celebrities, social events, etc, which enables a comprehensive understanding of sarcasm sentiment. Without this knowledge, it would be difficult for the model to accurately detect the true emotion of the post. By incorporating deep-level knowledge, the model can recognize that this tweet is related to the political situation and the author's frustration, thereby revealing its sarcastic emotion.

In this paper, we introduce a novel framework DeepMSD to incorporate underlying beneficial information for sarcasm reasoning, which leverages the ability of Large Vision-Language Models (LVLMs) to enhance the understanding of multimodal posts. Specifically, we first propose a Deep-level Knowledge Extraction Module (DKEM) to utilize the deep-level knowledge (induced from the LVLMs) to complement the existing instances with only surface-level knowledge. As shown in Figure 1, DKEM consists of three stages: 1) we first use

Received 13 June 2024; revised 23 October 2024 and 20 December 2024; accepted 11 January 2025. Date of publication 16 January 2025; date of current version 4 July 2025. This work was supported in part by the Youth Science Fund of the Natural Science Foundation of the Autonomous Region under Grant XQZX20240123 and in part by the National Natural Science Foundation of China under Grant U23B2053 and Grant 62176182. This article was recommended by Associate Editor D. Zhang. *(Corresponding author: Longbiao Wang.)*

Yiwei Wei is with the Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: wywy@tju.edu.cn).

Hengyang Zhou, Haitao Shi, and Zhiyang Jia are with the College of Petroleum Engineering, China University of Petroleum (Beijing) at Karamay, Karamay 834000, China (e-mail: hengyangzhou@outlook.com; shihaitao1111@gmail.com; jiazhiyang@cupk.edu.cn).

Shaozu Yuan, Meng Chen, and Xiaodong He are with JD AI Research, Beijing 101111, China (e-mail: yuanshaozu@jd.com; chenmengdx@gmail.com; hexiaodong@jd.com).

Longbiao Wang is with the Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, and also with Huiyan Technology (Tianjin) Company, Ltd., Tianjin 300350, China (e-mail: longbiao_wang@tju.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2025.3530436

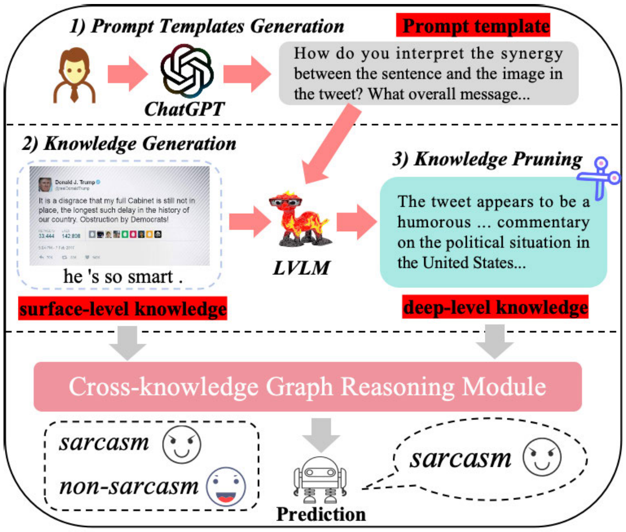

1) Prompt Templates Generation
Prompt template
ChatGPT
How do you interpret the synergy between the sentence and the image in the tweet? What overall message...
2) Knowledge Generation
Donald J. Trump
It is a disgrace that my full Cabinet is still not in place, the longest such delay in the history of our country. Obstruction by Democrats!
he 's so smart .
surface-level knowledge
LVLM
3) Knowledge Pruning
The tweet appears to be a humorous ... commentary on the political situation in the United States...
deep-level knowledge
Cross-knowledge Graph Reasoning Module
sarcasm
non-sarcasm
Prediction
sarcasm

Fig. 1. Here is an example of DeepMSD, which extracts underlying deep-level knowledge to facilitate multimodal sarcasm detection.

ChatGPT [18] to generate a series of prompt templates and select the best-performing one from the validation set "*How do you interpret the synergy between the sentence and the image in this tweet? What overall message do you think the tweeter is aiming to communicate?*"; 2) we then employ the large vision-language models (LVLMs, e.g., LLaVA [19]) to produce the underlying deep-level knowledge "*The tweet appears to be humorous …commentary on the political situation…*" based on the provided image and sentence; 3) due to noise and redundant context contained in the generated text, we further design a knowledge pruning method to eliminate these noise. Apart from DKEM, to emulate the way humans leverage prior knowledge to understand the sarcastic cues in multimodal posts, we introduce a Cross-knowledge Graph Reasoning Module (CGRM). This module establishes semantic relationships between deep-level knowledge and surface-level knowledge. Considering that deep-level knowledge relies on the understanding of the multimodal context, we exclude using visual information for graph construction to mitigate the impact of visual noise. CGRM constructs cross-knowledge graph representations based on textual semantic dependencies between deep-level knowledge and surface-level knowledge. This allows for the capture of incongruities using graph attention networks [20]. Subsequently, we employ the Transformer encoder to align and integrate cross-knowledge graph representations, thus obtaining the final prediction. In our experiments, DeepMSD achieves a new state-of-the-art performance on the public multimodal sarcasm detection dataset [21], yielding an accuracy score of 92.20.

To summarize, the main contributions of this paper are three-fold:

- In our limited knowledge, this is the first work leveraging underlying deep-level knowledge to augment the model's understanding in multimodal sarcasm detection task, inspired by the way humans understand sarcastic content.
- To achieve this, we have devised two novel modules: 1) a Deep-level Knowledge Extraction Module (DKEM), which utilizes the LVLM to produce Deep-level knowledge, and 2) a Cross-knowledge Graph Reasoning Module (CGRM), which emulates how humans use prior knowledge to understand the sarcastic cues in multimodal posts by establishing semantic relationships between different knowledge.
- Experiments conducted on the MSD benchmark demonstrate that our proposed model surpasses all previous models significantly, achieving an accuracy score of 92.20.

II. Related Work

A. Multi-Modal Sarcasm Detection

Multimodal Sarcasm Detection (MSD) is a substantial improvement beyond traditional text-based approaches by incorporating a diverse array of modalities, including images, to enhance classification accuracy. Driven by the escalating need to analyze multi-modal content proliferating on social media, numerous state-of-the-art models [9], [11], [12] have surfaced to address this complex challenge. These models effectively harness the complementary insights offered by different modalities, facilitating a more nuanced understanding of sarcasm within its diverse contextual frameworks. Notably, Schifanella et al. [22] spearheaded this endeavor by framing it as a multimodal classification task, successfully amalgamating manually crafted multimodal features. In the wake of this seminal work, various researchers [13], [21], [22] have introduced diverse multimodal networks, aiming to unearth the incongruities inherent in text and image data. Moreover, approaches like InCrossMGs [8], CMGCN [9], HKEmodel [10], MILNet [11] and GGSAM [16] have harnessed the power of transformers [23] and graph neural networks to model both intra-modality and inter-modality incongruities. Meanwhile, DIP [12] has made strides by designing a dual incongruity perceiving network to mine the sarcastic information from factual and affective levels. While current methods that utilize complex fusion techniques have demonstrated impressive results in MSD, they are inherently constrained by their reliance on surface-level information. As a result, they fail to effectively integrate deep-level knowledge.

B. Large Vision-Language Models

Recent advancements in Large Vision-Language Models (LVLMs) position them as key players in tackling diverse and general tasks, as highlighted in various studies [24], [25], [26], [27], [28]. Notably, [19] have made significant strides with the introduction of LLaVA. This model innovatively bridges the gap between the CLIP ViT-L/14 visual encoder [29] and the robust language model Vicuna [30] through a sleek projection matrix. Its two-stage instruction-tuning technique further enhances its versatility. Meanwhile, [27] proposed MMICL to tackle the complexities of multi-modal prompts from a fresh perspective. This method offers valuable insights for refining LVLMs, addressing both modeling and data-related

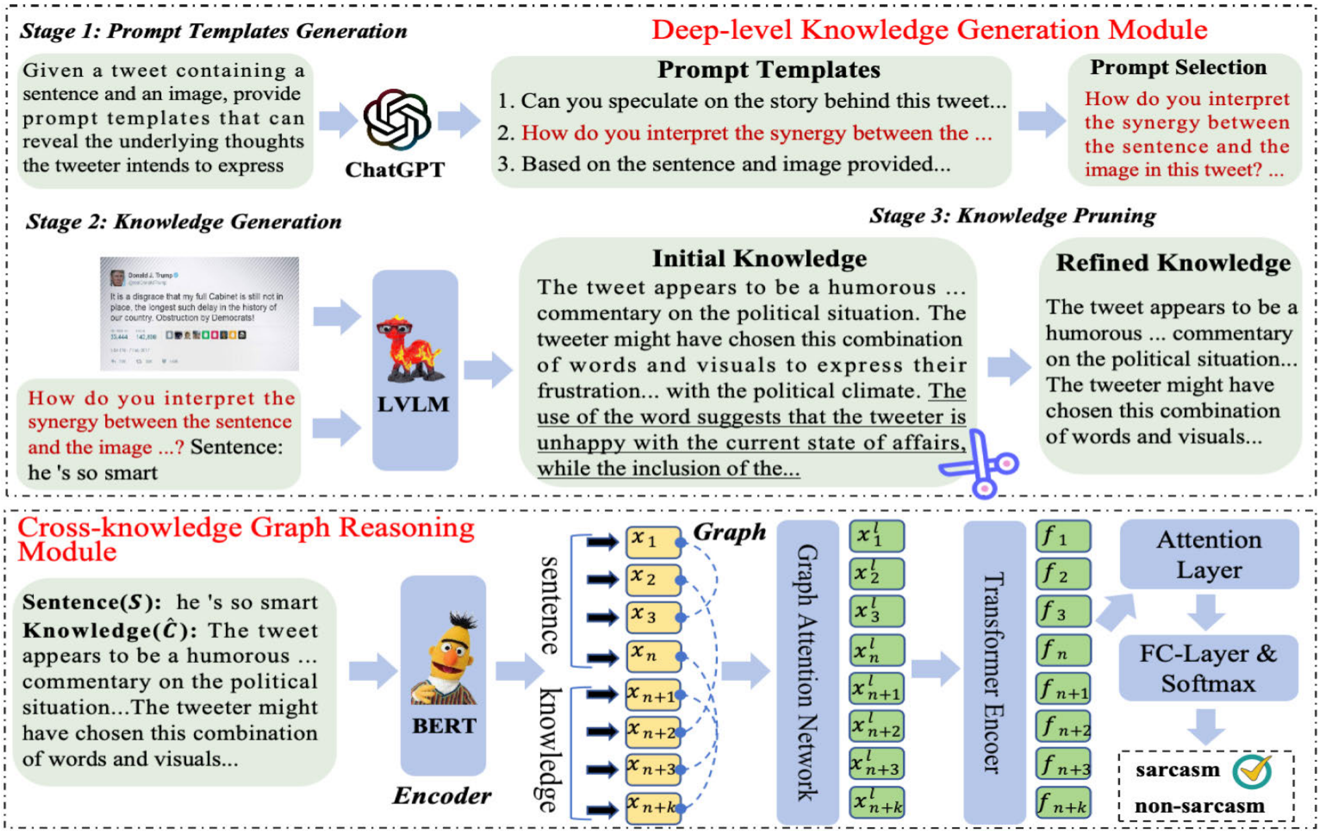

Stage 1: Prompt Templates Generation
Deep-level Knowledge Generation Module
Given a tweet containing a sentence and an image, provide prompt templates that can reveal the underlying thoughts the tweeter intends to express
ChatGPT
Prompt Templates
1. Can you speculate on the story behind this tweet...
2. How do you interpret the synergy between the ...
3. Based on the sentence and image provided...
Prompt Selection
How do you interpret the synergy between the sentence and the image in this tweet? ...
Stage 2: Knowledge Generation
Stage 3: Knowledge Pruning
Donald J. Trump
It is a disgrace that my full Cabinet is still not in place, the longest such delay in the history of our country. Obstruction by Democrats!
How do you interpret the synergy between the sentence and the image ...? Sentence: he 's so smart
LVLM
Initial Knowledge
The tweet appears to be a humorous ... commentary on the political situation. The tweeter might have chosen this combination of words and visuals to express their frustration... with the political climate. The use of the word suggests that the tweeter is unhappy with the current state of affairs, while the inclusion of the...
Refined Knowledge
The tweet appears to be a humorous ... commentary on the political situation... The tweeter might have chosen this combination of words and visuals...
Cross-knowledge Graph Reasoning Module
Sentence(S): he 's so smart
Knowledge(Ĉ): The tweet appears to be a humorous ... commentary on the political situation...The tweeter might have chosen this combination of words and visuals...
BERT
Encoder
sentence
knowledge
x1
x2
x3
xn
xn+1
xn+2
xn+3
xn+k
Graph
Graph Attention Network
x^l_1
x^l_2
x^l_3
x^l_n
x^l_n+1
x^l_n+2
x^l_n+3
x^l_n+k
Transformer Encoer
f1
f2
f3
fn
fn+1
fn+2
fn+3
fn+k
Attention Layer
FC-Layer & Softmax
sarcasm
non-sarcasm

Fig. 2. The overall architecture of the proposed DeepMSD. The Deep-level Knowledge Extraction Module (DKEM) depicted in the above box comprises three key stages. In Stage 1, ChatGPT is employed to offer prompt templates. Moving to Stage 2, LVLMs are required to generate knowledge utilizing these templates. Finally, in Stage 3, the deep-level knowledge generated is refined through a pruning method. Below, the Cross-knowledge Graph Reasoning Module (CGRM) is illustrated, encompassing three primary modules: Bert Encoder, Graph Attention network, and Transformer Encoder.

challenges. In a parallel development, [25] proposed mPUG-Owl2, a multi-modal language model designed for enhanced collaboration between modalities. Its modular architecture, centered around a versatile language decoder, underscores its potential for facilitating rich cross-modal interactions and advancing the field. Due to the advantages of LVLMs, we leverage these models to generate deep-level knowledge to facilitate multimodal sarcasm detection.

C. Knowledge Augmented Models

Leveraging external knowledge to enhance models has emerged as a promising approach in knowledge-intensive tasks [31], [32], [33], [34]. One approach involves retrieval-augmented LMs, which retrieve relevant passages and integrate them into LMs using techniques such as chunked cross-attention [35], dual instruction tuning [36], and self-reflection [37]. However, the potential of augmenting MSD, with underlying deep-level knowledge induced from LVLMs remains largely unexplored.

III. Methodology

In this section, we introduce the details of the components of DeepMSD, as depicted in Figure 2. Generally, DeepMSD is composed of a Deep-level Knowledge Extraction Module (DKEM), which is designed to extract deep-level knowledge utilizing large vision-language models. It also includes a Cross-knowledge Graph Reasoning Module (CGRM), which captures the incongruities between different knowledges by constructing a cross-knowledge graph based on semantic dependencies.

A. Deep-Level Knowledge Extraction Module

As illustrated in the upper dashed box of Figure 2, our deep-level knowledge extraction module consists of three stages: 1) we use large language models ChatGPT to produce prompt templates; 2) these prompt templates are fed into the LVLM with sentence S and image V to generate deep-level knowledge; 3) we further utilize pruning method to remove noise and redundant context contained in the generated deep-level knowledge, avoiding the interference of irrelevant information on the model.

1) Stage 1: Prompt Templates Generation: This stage aims to provide high-quality prompt templates that facilitate generating deep-level knowledge, enabling the LVLM to grasp the intention and background behind the tweet more effectively. To achieve that, we employ ChatGPT to produce a set of candidate prompt templates. Specifically, we begin by selecting a question: "*Given a tweet containing a sentence and an image, provide prompt templates that can reveal the underlying thoughts the tweeter intends to express*", and use the GPT-3.5 model to generate potential templates. Then, we incorporate a placeholder "Sentence: < x>" at the end of the prompt template to accommodate input text S of the tweet. Subsequently, we validate these prompts on the validation set and select the best one for deep-level content generation.

2) Stage 2: Knowledge Generation: In this stage, we utilize prompt templates in stage 1 to generate deep-level knowledge, by inputting the given image (V) and sentence (S) into LLaVA [19]. Since LLaVA's pretraining corpus encompasses a wide range of world facts and can understand various ironic cues, we do not further finetune LLaVA. Specifically, we formulate the prompt by substituting the placeholder "< x>" in the prompt template with the designated sentence. Furthermore, we introduce a special token "< image>" at the beginning of the instruction to denote image slots for the image V. Through this process, we obtain the initial deep-level knowledge C, which encapsulates the fundamental understanding extracted from both the image and the sentence.

3) Stage 3: Knowledge Pruning: While initial deep-level knowledge C can uncover latent sarcastic sentiment embedded within multimodal tweets more effectively compared to surface-level information, it still contains irrelevant information that may impact performance. As depicted in the raw knowledge of Figure 2, the first two sentences have already revealed the background and intent of the post, which are the most critical parts. Although the subsequent sentences indicated by underscores offer further details on the event background, they also contain an abundance of irrelevant information. This irrelevant information not only introduces unnecessary interference to the language modeling but also brings an extra computational burden. Therefore, we devised a knowledge pruning method to remove the irrelevant sentences retained in the initial knowledge based on a pruning threshold β. To ensure accurate sentence segmentation, we leverage the SpaCy model[1] to comprehend and divide the text into grammatically accurate sentences. Due to its superior sentence boundary identification capabilities, it can extract important and complete sentences, which is crucial for our pruning process. By leveraging the SpaCy model, we can extract the top β important sentences from the initial knowledge based on their semantic structures, discarding the irrelevant sentences. Thus, we obtain a refined knowledge, termed as $\hat{C}$.

B. Cross-Knowledge Graph Reasoning Module

To emulate the way humans leverage prior knowledge to understand the sarcastic cues in multimodal posts, we introduce a Cross-knowledge Graph Reasoning Module (CGRM). CGRM establishes the cross-knowledge graphs between deep-level knowledge ($\hat{C}$) and textual surface-level knowledge (S). And the cross-knowledge graphs are fused through the graph attention network and transformer encoder. Note that we exclude using visual information for graph construction. Because images contain numerous redundant and emotion-irrelevant elements, these can easily interfere with the model's understanding of the sentiment in the post. On the contrary, the deep-level knowledge produced by the proposed DKEM is more related to sentiment in the data.

The overall architecture of the proposed CGRM module is illustrated in the below box of Figure 2. Given the textual surface-level knowledge ($S = \{s_1, s_2, \ldots, s_n\}$) and the refined deep-level knowledge ($\hat{C} = \{c_1, c_2, \ldots, c_k\}$), we first concatenate them as $[S; \hat{C}]$. Then, we employ the pretrained BERT model [38] with an additional multi-layer perception to transform $[S; \hat{C}]$ into a sequence of token-level features $X = \{x_1, x_2, \ldots, x_{n+k}\}$, where $X \in \mathbb{R}^{(n+k)\times d}$. Subsequently, we construct cross-knowledge graphs by leveraging semantic dependencies among tokens in X to establish cross-knowledge relationships. To achieve that, we take tokens in X as graph nodes and utilize dependency relations between words extracted by SpaCy model as edges. During constructing cross-knowledge graphs, an edge is established when there is a dependency between two words. We then employ graph attention networks (GAT) [20] to model the cross-knowledge graphs. GAT utilizes the self-attention mechanism to distinguish the complex relationships among nodes within cross-knowledge graphs. This approach assigns different weights to the edges between nodes, effectively capturing the structural intricacies of the graph. Through this process, the model can focus on those emotion-related relationships, thereby better understanding the tweet content. Here, we illustrate the details of GAT to cross-knowledge graphs as follows:

$$\alpha_{ij}^{l} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\vec{a}_l^{T}\left[W_l\vec{x}_i^{l}\,\|\,W_l\vec{x}_j^{l}\right]\right)\right)}{\sum_{k\in\mathcal{N}(I)\cup i}\exp\left(\mathrm{LeakyReLU}\left(\vec{a}_l^{T}\left[W_l\vec{x}_i^{l}\,\|\,W_l\vec{x}_k^{l}\right]\right)\right)} \tag{1}$$

$$\vec{x}_i^{l+1} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij} W_l \vec{x}_j^{l}\right) \tag{2}$$

where W_l and $\vec{a}$ are the learnable parameter of the l-th GAT layer, $\vec{x}_i^{l}$ is the feature vector of node i in the l-th layer, with $x_i^0 = x_i$ initialized from the token-level features X. $\|$ denotes concatenation, $\mathcal{N}(i)$ includes node i and its neighbors, $LeakyReLU$ and σ are the activation functions. As such, the output of the L layer is $\hat{X} = \{x_1^L, x_2^L, \ldots, x_{n+k}^L\}$ with $\hat{X} \in \mathbb{R}^{(n+k)\times d}$.

Subsequently, we use a multi-layer Transformer-Encoder as a cross-knowledge graph fusion layer which will align and fuse the cross-knowledge graph representations $\hat{X}$. Then the fusion sequence features can be obtained. It is as follows:

$$F \in \mathbb{R}^{(n+k)\times d} = \{f_1, f_2, \ldots, f_{n+k}\} = TE(\hat{X}) \tag{3}$$

where F denotes the fused graph representation, TE denote the vanilla Transformer-Encoder and f indicate feature for each layer.

After acquiring the sequence features, to utilize them for the ultimate classification task, we employ a basic attention layer to obtain the final representation:

$$\tilde{r}_i = GELU(f_i W_1 + b_1) W_2 + b_2 \tag{4}$$

$$Q = GELU((\sum_{i=1}^{k+s} exp(\frac{\tilde{q}_i}{\sum_{j=1}^{n+k}\tilde{q}_j})(f_i))W_3 + b_3) \tag{5}$$

where $GELU$ is the activation function, $Q \in \mathbb{R}^d$.

[1] https://spacy.io/

TABLE I
STATISTICS OF THE EXPERIMENTAL DATA

Dataset	Label	Train	Val	Test
HFM	Positive	8642	959	959
	Negative	11174	1451	1450
	All	19816	2410	2409

C. Training Loss

As shown in the below dashed box of Figure 2, we feed the above representation Q into the fully connected layer and employ the softmax function for sarcasm detection. We use the cross-entropy loss as the classification loss and it is as follows:

$$\mathcal{L} = CrossEntropy(GELU(QW_{sc} + b_{sc})) \quad (6)$$

where W_{sc} and b_{sc} denotes the learnable parameters.

IV. EXPERIMENTS

A. Datasets

1) Multi-Modal Sarcasm Dataset: Our research utilized the publicly accessible multi-modal sarcasm detection dataset named Hierarchical Fusion Model (HFM) dataset [21], comprising tweets that are sarcasm (labeled as positive examples) and non-sarcasm (labeled as negative examples), along with associated images for each tweet. Following the division approach used by Cai et al. [21], we split the data into training, validation, and test sets with an 80%:10%:10% distribution, aiming for a straightforward comparison. The dataset's detail is presented in Table I.

B. Experimental Settings

For a fair comparison, we follow previous works [7], [8], [10], [21] to pre-process the dataset. We utilize the pre-trained BERT-base model to embed each word of the cross-knowledge information, setting the textual embedding size to 768. Subsequently, an additional MLP is employed atop BERT to obtain token-level features with dimensions of 200. The number of multi-head self-attention layers in the transformer encoder is set to 6. Within the knowledge pruning stage of the proposed DKEM module, the pruning threshold β is set to 2. For prompt template generation, we employ GPT-3.5-turbo to generate prompt templates. In the knowledge generation stage, we employ the LLaVA-v1.5 (13B) model [19] for deep-level knowledge generation. Additionally, we explore several other large vision-language models, the details of which will be discussed in the ablation studies. During the training stage, we utilize Adam as the optimizer with a learning rate of 2e-5. Furthermore, we set weight decay as 5e-3, batch size as 32, and dropout rate as 0.5 to train the model. To prevent overfitting, we implement early stopping with a patience of 5. Following [9], [21], we use Accuracy, Precision, Recall, and F1-score to measure the model performance. To mitigate the impact of imbalanced data distribution and comprehensively evaluate the model's performance, we also report macro-average scores in the main results.

C. Baseline Models

To fully validate the performance of DeepMSD, we select both unimodal and multi-modal baselines.

1) Unimodal Baselines: For text-based methods, we adopt TextCNN [40], Bi-LSTM [42], SIARN [1] which adopts inner-attention for text sarcasm detection, SMSD [3] which employs a self-matching network to capture incongruity information for sarcasm detection, and BERT [38] is a pre-trained model for text classification. For pure image-based methods, we employ the pooled feature of the pre-trained Resnet model [39] and the [CLS] token obtained by the pre-trained ViT model to detect sarcasm.

2) Multi-modal Baselines: In our comprehensive evaluation, we benchmark against a diverse array of multimodal methods to ensure a thorough comparative analysis. Among the notable approaches, **HFM** [21] introduces a hierarchical fusion model specifically tailored for multimodal sarcasm detection. **D&R** [7] constructs a decomposition and relation network to model the semantic association in the cross-modality context. **Att-BERT** [13] enhances this approach by incorporating sophisticated attention mechanisms aimed at sarcasm identification. **InCrossMGs** [8] leverages a heterogeneous graph structure, enabling the capture of ironic expressions from varied angles. **CMGCN** [9] innovates further by constructing a cross-modal graph for each data instance, thereby explicitly mapping the ironic relationships spanning across modalities. **HKEmodel** [10] advances a hierarchical analysis framework, exploring both atomic-level and composition-level congruences via graph neural networks for enhanced sarcasm detection. **multi-view CLIP** [41] leverages multi-grained cues from multiple perspectives for multi-modal sarcasm detection. **MILNet** [11] introduces a tripartite graph structure focused on identifying multimodal incongruities. **DIP** [12], representing the pinnacle of current advancements, adeptly models the incongruity at both factual and affective levels. **G^{2}SAM** [16] proposes a novel paradigm for handling multimodal sarcasm detection task by using graph-based global semantic awareness.

D. Main Results

To evaluate the proposed DeepMSD, we conduct a comparative analysis against state-of-the-art MSD methods using the HFM dataset, with results reported in Table II. From these results, we can draw different conclusions. Firstly, it is observed that text-based models exhibit superior performance compared to image-based models, as the text demonstrates more emotion-related information while visual data contains more irrelevant information. Furthermore, models that fuse text and image modalities surpass those reliant on a single modality, as the former approach offers additional cues for sarcasm detection, which emphasizes the significance of introducing different modalities to enhance the analysis. Nevertheless, merely depending on surface-level data from both different modalities restricted the ability to capture the underlying sentiment conveyed in tweets. Conversely, it can be seen that our proposed method significantly improves the performance of sarcasm detection for it introduces deep-level

TABLE II

EXPERIMENTAL RESULTS FOR MULTIMODAL SARCASM DETECTION. TO COMPREHENSIVELY EVALUATE OUR MODEL, WE ALSO REPORT THE METRICS IN TERMS OF MACRO-AVERAGE

MODALITY	METHOD	Ref.	Acc(%)	Pre(%)	Rec(%)	F1(%)	Macro-average		
							Pre(%)	Rec(%)	F1(%)
image	Resnet [39]	CVPR'16	64.76	54.41	70.8	61.53	60.12	73.08	65.97
	ViT [29]	ICLR'21	67.83	57.93	70.07	63.43	65.68	71.35	68.4
text	TextCNN [40]	EMNLP'14	80.03	74.29	76.39	75.32	78.03	78.28	78.15
	Bi-LSTM [9]	ACL'22	81.9	76.66	78.42	77.53	80.97	80.13	80.55
	SIARN [1]	ACL'18	80.57	75.55	75.7	75.63	80.34	78.81	79.57
	SMSD [3]	WWW'19	80.9	76.46	75.18	75.82	80.87	78.20	79.51
	BERT [38]	NAACL'19	83.85	78.72	82.27	80.22	81.31	80.87	81.09
multimodal	HFM [21]	ACL'19	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	D&R [7]	ACL'20	84.02	77.97	83.42	80.60	-	-	-
	Att-BERT [13]	EMNLP'20	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	InCrossMGs [8]	MM'21	86.1	81.38	84.36	82.84	85.39	85.8	85.6
	CMGCN [9]	ACL'22	86.54	-	-	82.73	-	-	-
	HKEmodel [10]	EMNLP'22	87.36	81.84	86.48	84.09	-	-	-
	Multi-view CLIP [41]	ACL'23	88.33	82.66	88.65	85.55	-	-	-
	MILNet [11]	AAAI'23	89.50	85.16	89.16	87.11	88.88	89.44	89.12
	DIP [12]	CVPR'23	89.59	87.76	86.58	87.17	88.46	89.13	89.01
	G^{2}SAM [16]	AAAI'24	90.48	87.95	89.02	88.48	89.44	89.79	89.65
	DeepMSD (Ours)	-	**92.20**	**91.79**	**88.85**	**90.30**	**91.15**	**93.57**	**92.34**

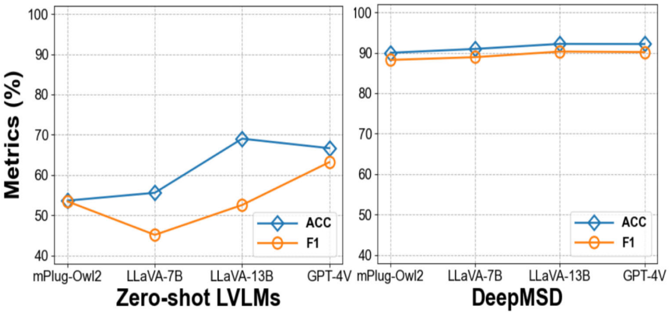
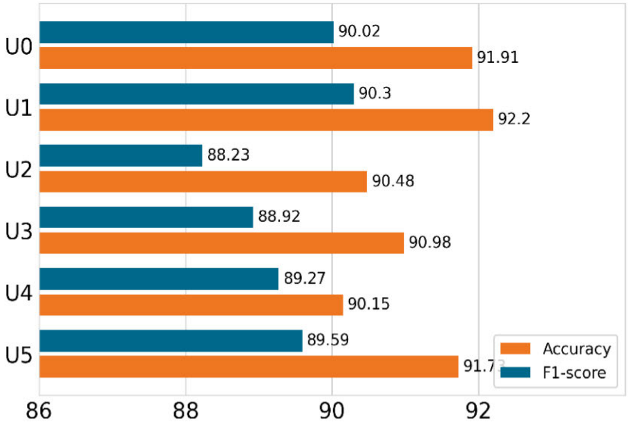


Fig. 3. Impact of employing various LVLMs. "Zero-shot LVLM" refers to using LVLM's zero-shot capability to directly predict the emotion behind the tweet. "DeepMSD" denotes our proposed model, which harnesses underlying deep-level LVLM-generated knowledge to facilitate multimodal sarcasm detection.



U0: Can you speculate on the story behind this tweet? What do you think motivated the tweeter to share this particular combination of words and visuals?

U1: How do you interpret the synergy between the sentence and the image in this tweet? What overall message do you think the tweeter is aiming to communicate?

U2: What do you think prompted the tweeter to pair this specific sentence with this particular image? What message do you believe they intend to send?

U3: Based on the sentence and image provided, what do you infer about the tweeter's perspective or opinion on the subject matter depicted?

U4: What is the direct correlation between the text and the image in the tweet, and how does this combination express a particular thought or sentiment?

U5: What could be the tweeter's intention or purpose behind pairing this specific image with the text?

Fig. 4. Effects of different types of prompt template. The orange and blue dash lines represent the accuracy and F1-score, respectively.

knowledge to boost the model's understanding of the sentiment behind the tweets. Overall, the proposed DeepMSD outperforms across all evaluated metrics, indicating the advantage of incorporating deep-level knowledge. Compared to the previous state-of-the-art work, G^{2}SAM [16], the accuracy has increased by 2%, and the F1 score has improved by 1%. This improvement is significant, considering that the improving margin of most previous state-of-the-art models was less than 1%, thereby illustrating the effectiveness of the proposed approach.

E. Ablation Study

In this section, we validate the effectiveness of the proposed DeepMSD through a series of distinct ablation studies.

1) Analyzing Effects of Different Knowledge: We initially delve into the impact of different types of knowledge on model efficacy. For simplicity, visual surface-level knowledge (image) is denoted as V, textual surface-level knowledge (text) as S, and the generated deep-level knowledge as $\hat{C}$. The ablation study results, which demonstrate the incremental

TABLE III
THE ABLATION RESULTS OF USING DIFFERENT KNOWLEDGE. HERE, VISUAL SURFACE-LEVEL KNOWLEDGE IS DENOTED AS V, TEXTUAL SURFACE-LEVEL KNOWLEDGE AS S, AND THE GENERATED DEEP-LEVEL KNOWLEDGE AS $\hat{C}$

Model	ACC(%)	Pre(%)	Rec(%)	F1(%)
Only V	70.15	62.83	75.37	68.53
Only $\hat{C}$	79.27	73.54	76.88	75.18
Only S	87.69	85.52	84.11	84.81
S+V	88.55	85.47	87.02	86.24
S+$\hat{C}$	**92.20**	91.79	**88.85**	**90.30**
S+V+$\hat{C}$	91.85	**92.39**	88.17	90.23

incorporation of these knowledge types, are detailed in Table III. The findings indicate that models that depend solely on visual knowledge (V) exhibit the poorest performance metrics, as there is less emotional information in the image. Interestingly, solely utilizing generated deep-level textual knowledge (S), the model achieves better performance compared to visual knowledge (V), indicating the rationality of introducing deep-level information. Given that deep-level information serves as a complement to textual data, text can more effectively convey emotions. Consequently, models relying on textual knowledge (S) achieved better results, and demonstrated superior performance, emphasizing the crucial role of textual data in the domain of multi-modal sarcasm detection. This observation is consistent with previous research conclusions [10]. However, the further introduction of visual knowledge (V) into this mix paradoxically harms the performance. In response to this phenomenon, we have identified two key factors. First, in multimodal sarcasm detection, the visual modality typically provides less salient information and introduces more redundant features compared to the text modality. Consequently, only a limited number of visual cues effectively convey sentiment, leading to an abundance of irrelevant visual elements and lower information density. In contrast, the text modality delivers richer, more direct sentiment cues due to its higher information density. This indicates that superficial knowledge embedded in visual data often acts as noise, potentially diminishing the model's overall accuracy. Second, since the extracted deep-level knowledge already incorporates the visual information from the images, reintroducing visual knowledge could lead to redundancy, impairing the model's performance.

2) Analyzing Effects of Different Prompt Templates: We then investigated the impact of different prompt templates on model performance, employing six distinct prompt templates labeled from U0 to U5. The experiential results are shown in Figure 4. It was observed that the context generated by different templates significantly affects the model's performance, aligning with the current academic discourse on the crucial role of prompt template selection for large language models. To achieve optimal performance, we selected U1 as the prompt template for our subsequent experiments.

3) Analyzing Impact of Different LVLMs: To investigate the impact of deep-level knowledge produced by different large vision-language models (LVLMs) on model performance, we selected three representative LVLMs for experimentation. We extracted deep-level knowledge using these models and systematically evaluated the contribution of each type of knowledge to model performance. Moreover, to confirm that the improvements of our proposed model are not directly correlated with the strong generalization abilities inherent in LVLMs (mPLUG-Owl2 [25], LLaVA [19], GPT-4V[2]), we conducted zero-shot performance evaluations of these models on the multimodal sarcasm dataset. The results of zero-shot LVLMs and our DeepMSD are all presented in Figure 3. It shows that our DeepMSD surpasses the zero-shot LVLMs, revealing that our improvement does not directly stem from the capabilities of the large models themselves, but from the underlying deep-level knowledge they provide. This is the motivation of this work, giving possible insights for incorporating large models into future work in the MSD field. Additionally, as shown in Fig. 3, the knowledge generated by different LVLMs also has an impact on the model's performance, which further demonstrates the differences in the reasoning capabilities of these models. Based on the experimental results, we select LLaVA-13B as our deep-level knowledge generation model.

TABLE IV
THE PERFORMANCE OF OTHER MODELS AFTER INTRODUCING DEEP-LEVEL KNOWLEDGE

Model	ACC(%)	Pre(%)	Rec(%)	F1(%)
multi-view CLIP	88.33	82.66	88.65	85.55
+ Knowledge	89.67	84.59	88.37	86.44
DIP	89.59	87.76	86.58	87.17
+ Knowledge	91.12	90.57	88.25	89.39
DeepMSD	**92.20**	**91.79**	**88.85**	**90.30**

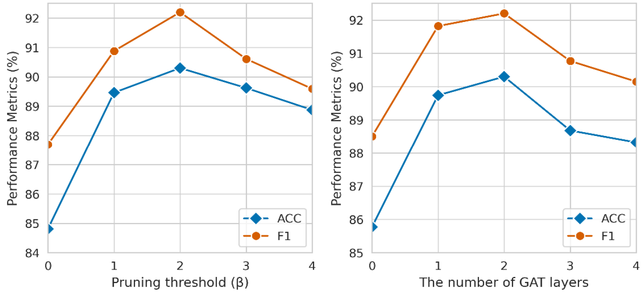


Fig. 5. The performance curve for multi-modal sarcasm detection with different settings.

F. Discussion

1) Universal Discussion: To verify the universality of our method, we also evaluate the performance of introducing deep-level knowledge on other MSD models. Given that G^{2}SAM [16] uses KNN for similarity prediction and has not been open-sourced, we conduct experiments using multi-view CLIP [41] and DIP [12]. Specifically, we directly incorporate the deep-level knowledge into the text data of each sample and train the models with the augmented samples. As shown in Table IV, the model's performance has also been improved

[2] https://cdn.openai.com/contributions/gpt-4v.pdf

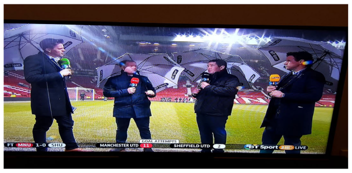
Sentence	**if only old trafford had some sort of studio facility to keep pundits dry !**	**Lable: Sarcasm**
Image		***Prompt U1*** **How do you interpret the synergy between the sentence and the image in this tweet? What overall message do you think the tweeter is aiming to communicate? Sentence: <X>**
LLaVA-13B	**The tweet is a playful commentary on the situation where the three men are standing on a field holding umbrellas, likely due to rain or wet conditions. *The tweeter is suggesting that if there were a studio facility at the location, the pundits would be protected from the rain and could continue their discussion without the need for umbrellas.* The overall message is a lighthearted and humorous take on the situation, *highlighting the inconvenience of the weather on the pundits' activities.***	
GPT4V	**Old Trafford, a world-renowned football stadium, presumably has the resources to provide proper facilities, yet the pundits are seen standing in the rain with umbrellas. The overall message the tweeter is aiming to communicate is a playful critique of the situation, *highlighting the absurdity of not having a covered studio area for the commentators to stay dry during inclement weather.* It underscores the expectation that a high-profile venue like Old Trafford should have better accommodations for its media personnel.**	
mPlug-Owl2	**The tweet is poking fun at the weather conditions at Old Trafford, a football stadium, during a match. *It implies that the weather is not ideal, as the pundits are holding umbrellas in the studio to avoid getting wet.* The overall message the tweeter is communicating is that the weather is not pleasant enough to enjoy the sporting event as intended, and it has to be managed with umbrellas to protect the pundits from the rain.**	

Fig. 6. Case study of different LVLMs for deep-level knowledge generation. Clues indicative of sarcastic content are highlighted in red within the deep-level knowledge.

TABLE V

THE PROMPT FOR ZERO-SHOT GENERATION. IN THIS PROMPT TEMPLATE, < IMAGE> REPRESENTS THE IMAGE DATA, < X> IS THE TEXT DATA

Zero-shot prompt
USER: <image>\n Examine the provided Twitter post, comprising both textual content and an image, to discern whether the tone conveys sarcasm. If it is sarcastic, answer 1, otherwise answer 0. Sentence:<x>.

after introducing deep-level knowledge, which suggests that deep-level knowledge is particularly crucial for multimodal sarcasm detection. In addition, DeepMSD achieves the best performance compared with the other models. This is mainly because our proposed CGRM module in DeepMSD can better capture sarcasm clues by modeling the semantic relationships between deep-level and surface-level knowledge.

2) Parameter Discussion: In Figure 5, we investigate the impact of the pruning threshold β in the knowledge pruning stage and the depth of the Graph Attention Network (GAT) in the cross-knowledge graph modeling module. The β threshold is crucial for trimming redundant information while retaining the initial β sentences of essential knowledge. We observed when β is set as 2, the model achieves optimal performance. Furthermore, the research delves into the depth of the Graph Attention Network (GAT), revealing that a two-layer architecture is most effective in capturing intricate relationships. Layers of Graph Attention Network (GAT) less than 2 layers will restrain the model's capacity while exceeding 2 layers can lead to overfitting and harm performance.

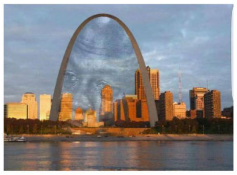
Sentence	tough day for st . louis . i feel so bad for them
Image	
Knowledge	The tweet featuring a sad face superimposed on the St. Louis Arch might be a form of satire or humor, possibly in response to a recent event or situation in St. The tweeter could be expressing empathy or concern for the city's residents, using the image of the sad face to evoke emotions and provoke thought.
Label	Sarcasm
Baseline	Non-sarcasm ❎
DeepMSD	Sarcasm ✅

Fig. 7. Case study of multimodal sarcasm detection. Clues indicative of sarcastic content are highlighted in red within the deep-level knowledge.

G. Case Study

In Figure 7, we present a case study illustrating the comparative performance between the baseline model, which relies solely on surface-level knowledge for sarcasm detection, and the proposed DeepMSD model that integrates deep-level knowledge to enhance sarcasm detection. This case reveals the inherent challenges models face in extracting sarcasm cues from content, as humans understand multimodal sarcasm is built upon an associating underlying deep-level knowledge. This deep-level knowledge necessitates broad common sense support, posing a challenge for models trained solely on sarcasm datasets to comprehend. By leveraging large vision-language models' common sense, we can uncover the deep-level information behind the posts. As shown in the

TABLE VI
THE PROMPT TEMPLATE FOR DEEP-LEVEL KNOWLEDGE GENERATION

U0
USER: <image>\n Can you speculate on the story behind this tweet? What do you think motivated the tweeter to share this particular combination of words and visuals? Sentence:<x>.
U1
USER: <image>\n How do you interpret the synergy between the sentence and the image in this tweet? What overall message do you think the tweeter is aiming to communicate? Sentence:<x>.
U2
USER: <image>\n What do you think prompted the tweeter to pair this specific sentence with this particular image? What message do you believe they intend to send? Sentence:<x>.
U3
USER: <image>\n Based on the sentence and image provided, what do you infer about the tweeter's perspective or opinion on the subject matter depicted? Sentence:<x>.
U4
USER: <image>\n What is the direct correlation between the text and the image in the tweet, and how does this combination express a particular thought or sentiment? Sentence:<x>.
U5
USER: <image>\n What could be the tweeter's intention or purpose behind pairing this specific image with the text? Sentence:<x>.

example, the presence of this deep-level information significantly assists the model in capturing these ironic cues. Therefore, this results in a more accurate prediction of the case's sarcastic emotion, highlighting the effectiveness of incorporating deep contextual understanding in sarcasm detection tasks.

In Figure 6, we also analyze the deep-level knowledge generated by three LVLMs (mPLUG-Owl2, LLaVA-13B, GPT-4V) through case study. The results show that the deep-level knowledge generated by all three models clearly reflects the deep sentiment embedded in the samples, which corroborates the findings in Fig. 2, where the deep-level knowledge generated by these models significantly improves MSD. Further analysis reveals that both the LLaVA-13B and GPT-4V explicitly highlight the tweeter's dissatisfaction with the lack of a covered studio for the commentators, while the mPLUG-Owl2 model fails to do so. This aligns with the results in Figure 3(b), where the performance of LLaVA-13B and GPT-4V is similar, while mPLUG-Owl2 performs worse, further confirming the differences in performance across different large models in MSD tasks.

H. Prompt Template

Table V presents a detailed list of prompts utilized for implementing the zero-shot generation, which is introduced in Section 5.2 of the main text. Due to the lack of instruction-follow data in the multimodal sarcasm detection task, we carefully design an instruction template following the conversational format of the LLaVA, as shown in table VII.

TABLE VII
INSTRUCTION TEMPLATE FORMAT OF THE LLAVA

###USER: <image>[prompt templates]<text>
###ASSISTANT: <label>

In this prompt, < image> represents the image data, <x> is the text data, and < label> is the predicted label. The human prompt adopts the choice question-answering format, as shown in Table V.

In Table VI, we also present the prompt templates for deep-level knowledge generation. U0 to U5 denote five different prompt templates, respectively.

V. LIMITATION

Considering that directly incorporating the visual modality may affect the model's performance, it is worth continually exploring how to overcome this issue. In this paper, we propose a Cross-knowledge Graph Reasoning Module to integrate different modalities in representation aspects, which is an implicit way to build the relations of fine-grained features, such as visual objects, and textual words. Previous work [43], [44] has proven that the novel CapsNets [45] and their part-whole relationships show advantages in modeling the relations among different elements by sparse modeling. Inspired by these works, we argue that explicitly introducing the relationship of fine-grained features via CapsNets can better guide the model to eliminate redundant features in visual knowledge. Thus it can further narrow the modalities gap and facilitate fusion for multimodal content understanding.

VI. CONCLUSION AND FUTURE WORK

This is the first work leveraging underlying deep-level knowledge to augment the model's understanding in multimodal sarcasm detection tasks. We designed two novel modules, a Deep-level Knowledge Extraction Module (DKEM) to produce Deep-level knowledge, and a Cross-knowledge Graph Reasoning (CGRM) to emulate how humans use prior knowledge to understand the sarcastic cues in multimodal posts by establishing semantic relationships between different knowledge. The experiments show our model achieves a new state-of-the-art performance on the public multimodal sarcasm detection dataset, with an accuracy score of 92.20. Our work validates the application of LVLM in the MSD domain is promising, and deserves further exploration. Considering that the generated deep-level knowledge relies entirely on unsupervised methods, we will explore ways to further enhance its reliability in future work. Additionally, we will consider incorporating CapsNets to learn multimodal relationships and improve the model's performance.

REFERENCES

[1] Y. Tay, A. T. Luu, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1010–1020.

[2] A. Ghosh and T. Veale, “Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 482–491.

[3] T. Xiong, P. Zhang, H. Zhu, and Y. Yang, “Sarcasm detection with self-matching networks and low-rank bilinear pooling,” in *Proc. World Wide Web Conf.*, May 2019, pp. 2115–2124.

[4] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, “Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3192–3203, May 2024.

[5] M. Ren, X. Huang, J. Liu, M. Liu, X. Li, and A.-A. Liu, “MALN: Multimodal adversarial learning network for conversational emotion recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6965–6980, Nov. 2023.

[6] C. Chen, X. Sun, Z. Tu, and M. Wang, “AST-GCN: Augmented spatial temporal graph convolutional neural network for gait emotion recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4581–4595, Jun. 2024.

[7] N. Xu, Z. Zeng, and W. Mao, “Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3777–3786.

[8] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, “Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs,” in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4707–4715.

[9] B. Liang et al., “Multi-modal sarcasm detection via cross-modal graph convolutional network,” in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist.*, 2022, pp. 1767–1777.

[10] H. Liu, W. Wang, and H. Li, “Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement,” 2022, *arXiv:2210.03501*.

[11] Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, and L. Nie, “Mutual-enhanced incongruity learning network for multi-modal sarcasm detection,” in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 8, pp. 9507–9515.

[12] C. Wen, G. Jia, and J. Yang, “DIP: Dual incongruity perceiving network for sarcasm detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2540–2550.

[13] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, “Modeling intra and inter-modality incongruity for multi-modal sarcasm detection,” in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 1383–1392.

[14] X. Wang, X. Sun, T. Yang, and H. Wang, “Building a bridge: A method for image-text sarcasm detection without pretraining on image-text data,” in *Proc. 1st Int. Workshop Natural Lang. Process. Beyond Text*, 2020, pp. 19–29.

[15] J. Zhang, X. Liu, Z. Wang, and H. Yang, “Graph-based object semantic refinement for visual emotion recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3036–3049, May 2022.

[16] Y. Wei, S. Yuan, H. Zhou, L. Wang, Z. Yan, R. Yang, and M. Chen, “G^{2}sam: Graph-based global semantic awareness method for multimodal sarcasm detection,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 9151–9159.

[17] W. X. Zhao et al., “A survey of large language models,” 2023, *arXiv:2303.18223*.

[18] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, 2022, pp. 27730–27744.

[19] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26286–26296.

[20] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” 2017, *arXiv:1710.10903*.

[21] Y. Cai, H. Cai, and X. Wan, “Multi-modal sarcasm detection in Twitter with hierarchical fusion model,” in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, 2019, pp. 2506–2515.

[22] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, “Detecting sarcasm in multimodal social platforms,” in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1136–1145.

[23] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[24] C. Fu et al., “MME: A comprehensive evaluation benchmark for multimodal large language models,” 2023, *arXiv:2306.13394*.

[25] Q. Ye et al., “mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration,” 2023, *arXiv:2311.04257*.

[26] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *Proc. 37th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates, 2023, pp. 34892–34916, Art. no. 1516.

[27] H. Zhao et al., “MMICL: Empowering vision-language model with multi-modal in-context learning,” 2023, *arXiv:2309.07915*.

[28] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, “Shikra: Unleashing multimodal LLM’s referential dialogue magic,” 2023, *arXiv:2306.15195*.

[29] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.

[30] W.-L. Chiang et al. (2023). *Vicuna: An Open-source Chatbot Impressing GPT-4 With 90%* ChatGPT Quality*. Accessed: Apr. 14, 2023. [Online]. Available: https://vicuna.lmsys.org

[31] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, “Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 48573–48602.

[32] C. Li, G. Cheng, and J. Han, “Boosting knowledge distillation via intra-class logit distribution smoothing,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4190–4201, Jun. 2024.

[33] H. Shi, M. Liu, X. Mu, X. Song, Y. Hu, and L. Nie, “Breaking through the noisy correspondence: A robust model for image-text matching,” *ACM Trans. Inf. Syst.*, vol. 42, no. 6, pp. 1–26, Nov. 2024.

[34] W. Wang, L. Ding, L. Shen, Y. Luo, H. Hu, and D. Tao, “WisdoM: Improving multimodal sentiment analysis by fusing contextual world knowledge,” 2024, *arXiv:2401.06659*.

[35] S. Borgeaud et al., “Improving language models by retrieving from trillions of tokens,” in *Proc. 39th Int. Conf. Mach. Learn.*, 2022, pp. 2206–2240.

[36] X. V. Lin et al., “RA-DIT: Retrieval-augmented dual instruction tuning,” 2023, *arXiv:2310.01352*.

[37] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-RAG: Learning to retrieve, generate, and critique through self-reflection,” 2023, *arXiv:2310.11511*.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*.

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] Y. Chen, “Convolutional neural network for sentence classification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1746–1751.

[41] L. Qin et al., “MMSD2.0: Towards a reliable multi-modal sarcasm detection system,” in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2023, pp. 10834–10845. [Online]. Available: https://aclanthology.org/2023.findings-acl.689

[42] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, 2005.

[43] Y. Liu, D. Cheng, D. Zhang, S. Xu, and J. Han, “Capsule networks with residual pose routing,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 8, 2024, doi: 10.1109/TNNLS.2023.3347722.

[44] Y. Liu, L. Zhou, G. Wu, S. Xu, and J. Han, “TCGNet: Type-correlation guidance for salient object detection,” *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6633–6644, Jul. 2024.

[45] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3859–3869.

Yiwei Wei is currently pursuing the Ph.D. degree with the School of Intelligence and Computing, Tianjin University. He is a Faculty Member with China University of Petroleum (Beijing) at Karamay. He has published over 15 papers in ACL, AAAI, ICASSP, and other venues. His research interests include multimodal sentiment analysis, natural language processing, and multimodal information extraction.

Hengyang Zhou is currently pursuing the B.E. degree in data science and big data technology with China University of Petroleum (Beijing) at Karamay. His current research interests include multimodal sentiment analysis and large language models.

Shaozu Yuan is currently an Algorithm Engineer at JD AI Research, China. His current research focuses on multimodal emotion analysis, natural language processing, and large language models. He has served as a reviewer for top-tier conferences including ACL, NeurIPS, and ACM MM.

Meng Chen is the AI Director of JD AI Research. Before that, he was a Research Scientist of Nuance Communications. He has published over 40 papers in prestigious academic conferences such as AAAI, IJCAI, ACM Multimedia, ACL, NAACL, ICASSP, Interspeech, and CIKM. His research interests include natural language processing, speech recognition, and dialogue systems. He also serves as a Program Committee member for several top-tier conferences including AAAI, ACL, EMNLP, NAACL, EACL, ICASSP, Interspeech, and ACM Multimedia. In 2023, he was honored with the Wuwenjun Artificial Intelligence Science and Technology Progress Award. He also received the Best Demo Award at ACM Multimedia in 2021.

Haitao Shi received the bachelor's and master's degrees from China University of Petroleum (East China). He is currently pursuing the Ph.D. degree with Shandong University. His research interests lie primarily in multimedia content analysis and information retrieval.

Zhiyang Jia received the M.S. degree in computer software and theory from Yunnan Normal University, Yunnan, China, in 2006. He is currently a Full Professor with the Department of Computer Science, China University of Petroleum (Beijing) at Karamay, Xinjiang, China. His current research interests include the IIoT, edge computing, and prognostics and health management.

Longbiao Wang (Member, IEEE) received the Dr.-Eng. degree from Toyohashi University of Technology, Toyohashi, Japan, in 2008. From 2008 to 2012, he was an Assistant Professor with the Faculty of Engineering, Shizuoka University, Shizuoka, Japan. From 2012 to 2016, he was an Associate Professor with Nagaoka University of Technology, Nagaoka, Japan. He is currently a Professor, the Director of Tianjin Key Laboratory of Cognitive Computing and Application, and the Vice Dean of the School of Artificial Intelligence, Tianjin University, Tianjin, China. His research interests include robust speech recognition, speaker recognition, acoustic signal processing, and natural language processing.

Xiaodong He (Fellow, IEEE) received the bachelor's degree from Tsinghua University, Beijing, in 1996, the M.S. degree from the Chinese Academy of Sciences, Beijing, in 1999, and the Ph.D. degree from the University of Missouri–Columbia in 2003. He is the Vice President of JD.com and the Director of JD AI Research. He is also an Affiliate Professor with the University of Washington, Seattle, serves in doctoral supervisory committees. Before joining JD.com, he was with Microsoft, for about 15 years, as the Principal Researcher and the Research Manager of DLTC with Microsoft Research, Redmond, WA, USA. He has published more than 200 papers in ACL, EMNLP, NAACL, CVPR, SIGIR, WWW, CIKM, NIPS, ICLR, ICASSP, PROCEEDINGS OF THE IEEE, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, *IEEE Signal Processing Magazine*, and other venues. His research interests are mainly in artificial intelligence areas, including deep learning, natural language, computer vision, speech, information retrieval, and knowledge representation. He is a CAAI Fellow. He received several awards, including the Outstanding Paper Award at ACL 2015.