

# 000 COMPLETE MULTI-MODAL METRIC LEARNING FOR 001 002 MULTI-MODAL SARCASM DETECTION 003

004 **Anonymous authors**

005 Paper under double-blind review

## 006 ABSTRACT

011 Multi-modal sarcasm detection identifies sarcasm from text-image pairs, an essen-  
012 tial technology for accurately understanding the user’s real attitude. Most research  
013 extracted the incongruity of text-image pairs as sarcasm information. However,  
014 these methods neglected inter-modal or intra-modal incongruities in fact and sen-  
015 timent perspectives, leading to incomplete sarcasm information and biased per-  
016 formance. To address the above issues, this paper proposes a complete multi-  
017 modal metric learning network (CMMML-Net) for multi-modal sarcasm detection  
018 tasks. Specifically, CMMML-Net utilizes a fact-sentiment multi-task representation  
019 learning module to produce refined fact and sentiment text-image representation  
020 pairs. It then designs a complete multi-modal metric learning to iteratively calcu-  
021 late inter-modal and intra-modal incongruities in fact and sentiment metric spaces  
022 (e.g., fact and sentiment metric space), efficiently capturing complete multi-modal  
023 incongruities. CMMML-Net performs well in explicitly capturing comprehensive  
024 sarcasm information and obtaining discriminative performance via deep metric  
025 learning. The state-of-the-art performance on the widely-used dataset  
026 demonstrates CMMML-Net’s effectiveness in multi-modal sarcasm detection.

## 027 1 INTRODUCTION

030 Sarcasm is a widely used implicit expression in which the real attitude conflicts with the literal  
031 meaning (Gibbs, 1986). This incongruity between real attitude and literal meaning is a crucial clue  
032 for identifying sarcastic intent (Joshi et al., 2015). Multi-modal sarcasm detection captures the  
033 user’s real attitude by identifying incongruities across and within various modalities. It has a wide  
034 range of applications in social media monitoring and management, intelligent interactive systems,  
035 news dissemination, and public opinion analysis, etc. With the development of social media, it has  
036 garnered significant attention (Kolchinski & Potts, 2018; Desai et al., 2022).

037 Multi-modal sarcasm detection captures multi-modal incongruity from a fact perspective or senti-  
038 ment perspective. Fact incongruity means sarcasm occurs when the literal meaning and the observed  
039 facts unexpectedly contrast (Grice, 1978; McDonald, 1999). Sentiment incongruity means sarcasm  
040 often occurs when the literal meaning is positive while the observed emotion is negative (Sperber &  
041 Wilson, 1987). Most research focuses on fact incongruity or sentiment incongruity. For example,  
042 Yue et al. (2023) leveraged prior knowledge from ConceptNet and contrastive learning to improve  
043 factual semantic incongruities in sarcasm detection tasks. Liang et al. (2022) constructed cross-  
044 modal graphs to extract sentiment incongruities and predict sarcasm. However, the lack of either the  
045 fact or sentiment perspective weakens sarcasm detection performance, as demonstrated in Figures 1  
046 (a) and (b). Therefore, this paper focuses on mining complete incongruity features, including fact  
047 incongruity and sentiment incongruity.

048 Additionally, most multi-modal sarcasm detection research only focused on inter-modal incongruity  
049 as sarcasm information. For instance, Cai et al. (2019) and Xu et al. (2020) utilized cross-modal  
050 attention mechanisms and fusion strategies to discover factual semantic relevance in inter-modal  
051 contexts. Wang et al. (2024b) designed an align-fuse-collaborate mechanism to enhance the inter-  
052 modal incongruities in sarcasm information fusion. However, these works neglected the importance  
053 of intra-modal incongruity in sarcasm detection, leading to incomplete incongruities and biased per-  
054 formance. Moreover, intra-modal incongruity is more effective and provides supplementary infor-  
055 mation in some sarcastic scenarios, such as when the information between modalities is congruous

or when one of the modalities has too little information, as shown in the Figure 1 (c) and (d). Thus, this work focuses on an efficient capture strategy for complete multi-modal incongruities.



(a) Fact incongruity

(b) Sentiment incongruity

(c) Intra text incongruity

(d) Intra image incongruity

	(a)	(b)	(c)	(d)
HFM	✓	✗	✗	✗
CMGCN	✗	✓	✗	✗
Ours	✓	✓	✓	✓

Figure 1: The test results of CMML-Net and different models on several examples including fact incongruity, sentiment incongruity, intra-modal text incongruity, and intra-modal image incongruity. HFM (Cai et al., 2019) utilized inter-modal attention to extra fact incongruity. CMGCN(Liang et al., 2022) built cross-modal graphs to extract sentiment incongruity.

This paper proposes a complete multi-modal metric learning network (CMML-Net) to efficiently capture complete multi-modal incongruity for sarcasm detection tasks. Specifically, CMML-Net utilizes a fact-sentiment multi-task representation learning module to extract fine-grained fact representations and sentiment representations from text-image data via Yolo-task and SenticNet-task. It then presents a fact-sentiment dual-stream network to construct fact and sentiment metric spaces for the following comprehensive incongruity capture. In the metric spaces, it designs complete multi-modal metric learning to iteratively calculate inter-modal and intra-modal incongruities in fact and sentiment metric spaces **a unified space (e.g., fact and sentiment metric space)**. CMML-Net efficiently and explicitly captures complete multi-modal incongruities as effective sarcasm information and obtains comprehensive performance for multi-modal sarcasm detection. The state-of-the-art performance on the widely-used dataset demonstrates the superiority of CMML-Net in multi-modal sarcasm detection tasks. We released the codes and parameters to facilitate the research community. [<https://anonymous.4open.science/r/CMML-Net.>]

The main contributions of our paper can be summarized as follows:

- To our knowledge, the CMML-Net is the first work in multi-modal sarcasm detection to introduce deep metric learning to iteratively and explicitly calculate complete multi-modal incongruities in fact and sentiment perspectives **inter-modal and intra-modal incongruities in a unified space (e.g., fact and sentiment metric space)**. It efficiently handles the biased performance of sarcasm detection models through comprehensive sarcasm information.
- CMML-Net is an effective fact-sentiment dual-stream framework in multi-modal sarcasm detection tasks. It contains a fact-sentiment multi-task representation learning module and fact-sentiment dual-stream network, gradually capturing complete multi-modal incongruities in fact and sentiment perspectives.

## 2 RELATED WORK

### 2.1 SINGLE-MODAL SARCASM DETECTION

Early research in sarcasm detection focused on text-based methods (Zhang et al., 2016; González-Ibáñez et al., 2011; Riloff et al., 2013). SIARN (Tay et al., 2018) introduces an attention-based model to capture contrast and incongruity. SMSD (Xiong et al., 2019) utilizes a self-matching

network to explore word interactions. As research expanded, image-based sarcasm detection also gained attention. Studies on memes revealed that images alone could convey sarcasm, even without any textual content(Sharma et al., 2020; Maity et al., 2022). Certain visual features in the images play a critical role in expressing this sarcasm.

## 2.2 MULTI-MODAL SARCASM DETECTION

Multi-modal sarcasm detection has primarily focused on the image-text modality. Early works employed inter-modal attention to explore associations between modalities. For instance, HFM (Cai et al., 2019) utilized inter-modal attention to guide hierarchical modality fusion, while D&RNet (Xu et al., 2020) modeled semantic associations between images and text using inter-modal attention to capture commonalities and differences. Similarly, CMGCN (Liang et al., 2022) built cross-modal graphs to detect sarcasm by extracting sentiment incongruities. However, these works primarily emphasize inter-modal incongruities, leaving intra-modal incongruities underexplored.

Some studies have recognized the importance of intra-modal incongruities. For instance, Att-Bert (Pan et al., 2020) and Multi-View CLIP (Qin et al., 2023) employed separate attention mechanisms, and InCrossMGs (Liang et al., 2021) used distinct GCNs to extract inter-modal and intra-modal incongruities separately. However, they focus only on factual semantic incongruities. Moreover, DMSD-CL Jia et al. (2024) and G2SAM(Wei et al., 2024) learn sarcasm patterns through sample-level differences.

In addition to these models, more nuanced incongruity detection frameworks have emerged. DIP (Wen et al., 2023) proposed a dual-stream architecture with semantic and sentiment streams to identify inter-modal incongruities. FSICN (Lu et al., 2024) introduced a fact-sentiment dual-stream network to capture incongruities. Another recent work, KnowleNet (Yue et al., 2023), improved factual incongruity detection using prior knowledge from ConceptNet and contrastive learning. In contrast, MuMu (Wang et al., 2024b) designed an align-fuse-collaborate mechanism to enhance inter-modal incongruity detection. Despite these advancements, these models remain focused on inter-modal incongruities.

CMMML-Net, our proposed model, addresses these gaps by explicitly and efficiently capturing complete multi-modal incongruities in fact and sentiment perspectives via deep metric learning.

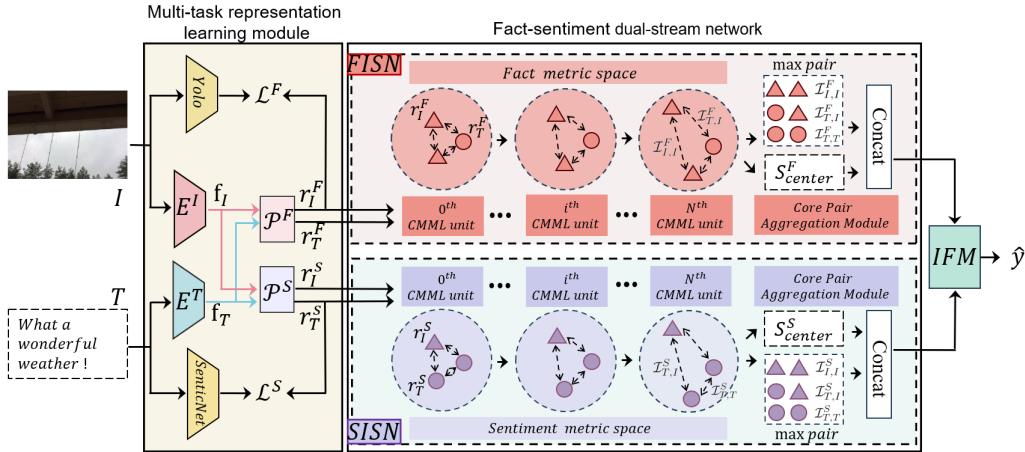


Figure 2: The overall architecture of CMMML-Net.

## 3 PROPOSED METHOD

There are many sarcastic situations in multi-modal sarcasm detection tasks, which can be easily missed and lead to biased recognition performance. Therefore, we propose a complete multi-modal metric learning network to efficiently capture complete multi-modal incongruity for multi-modal sarcasm detection tasks. The framework of CMMML-Net is shown in Figure 2. CMMML-Net comprises

162 a fact-sentiment multi-task representation learning module, a fact-sentiment dual-stream network,  
 163 and an incongruity fusion module (IFM). The details of each module are as follows.  
 164

165 Multi-modal sarcasm detection aims to determine whether a given text-image pair conveys sarcasm.  
 166 for a dataset  $D$  consisting of multi-modal samples, each sample  $d \in D$  includes a sentence  $T$  with  
 167  $n$  words  $\{t_1, t_2, t_3, \dots, t_n\}$  and an image  $I$  divided into  $m$  patches  $\{v_1, v_2, v_3, \dots, v_m\}$ .  
 168

### 3.1 FACT-SENTIMENT MULTI-TASK REPRESENTATION LEARNING MODULE

170 The fact-sentiment multi-task representation learning module aims to produce refined fact and  
 171 sentiment text-image representation pairs for the following complete multi-modal incongruity capture.  
 172 It leverages multi-task learning to construct fact and sentiment representation spaces. It mainly con-  
 173 sists of fact text-image representation extraction and sentiment text-image representation extraction.  
 174

175 We utilize the pre-trained and frozen CLIP text and image encoders to obtain public task-agnostic  
 176 general features (Radford et al., 2021). Specifically, The text encoder maps  $T$  into word-level em-  
 177 beddings  $f_T = \text{CLIP}_{\text{text}}(T)$ , and the image encoder transforms  $I$  into patch-level embeddings  
 $f_I = \text{CLIP}_{\text{image}}(I)$ .  
 178

179 Fact text-image representation extraction obtains effective fact text-image representations via Yolo-  
 180 task. Yolo queries object existence vectors  $y^F \in [0, 1]^k$  from the input image  $I$ , where  $k$  is the  
 181 number of detected object categories and the superscript capital letter  $F$  stands for fact. The vectors  
 182 guide the fact projection layer  $\mathcal{P}^F$  to project the image features  $f_I$  into fact image representation  $r_I^F$ .  
 183 The shared fact projection layer similarly projects the text features  $f_T$  into fact text representation  
 $r_T^F$ . Both representations are implicitly aligned in a unified fact representation space  $\mathcal{S}^F$ , which  
 184 facilitates fine-grained factual semantic inference. The binary cross-entropy loss (BCE) optimizes  
 185 the projection by minimizing the difference between the predicted object existence vector  $\hat{y}^F$  and  
 186 the pseudo labels  $y^F$ :  
 187

$$\mathcal{L}^F = - \sum_{i=1}^k (\mathbf{y}_i^F \log \hat{\mathbf{y}}_i^F + (1 - \mathbf{y}_i^F) \log(1 - \hat{\mathbf{y}}_i^F)) \quad (1)$$

190 Sentiment text-image representation extraction obtains effective sentiment text-image representa-  
 191 tions via SenticNet-task. SenticNet queries the input text and generates a sentiment polarity vector  
 $\mathbf{y}^S \in [-1, 1]^n$ . Here,  $n$  denotes the number of tokens,  $y_i^S = 0$  indicates no sentiment match (Liang  
 193 et al., 2022), and the superscript capital letter  $S$  stands for sentiment. The vectors guide the senti-  
 194 ment projection layer  $\mathcal{P}^S$  to project the text features  $f_T$  into sentiment text representation  $r_T^S$ . The  
 195 shared sentiment projection layer similarly projects the image features  $f_I$  into sentiment image rep-  
 196 resentation  $r_I^S$ . Both representations are implicitly aligned in the unified sentiment representation  
 197 space  $\mathcal{S}^S$ . The mean squared error (MSE) loss minimizes the prediction error.  
 198

$$\mathcal{L}^S = \frac{1}{n} \sum_{i=1}^n (y_i^S - \hat{y}_i^S)^2 \quad (2)$$

### 3.2 FACT-SENTIMENT DUAL-STREAM NETWORK

204 The fact-sentiment dual-stream (FSDS) network aims to capture intra-modal and inter-modal in-  
 205 congruities of fact text-image representations and sentiment text-image representations in unified  
 206 metric spaces. This obtains complete multi-modal incongruity and comprehensive recognition per-  
 207 formance for sarcasm detection tasks. FSDS network presents fact incongruity sub-network (FISN)  
 208 and sentiment incongruity subnetwork (SISN) to construct fact and sentiment metric spaces. It  
 209 then designs  $N$  complete multi-modal metric learning units to iteratively and explicitly calculate  
 210 complete multi-modal incongruities in fact and sentiment metric spaces a unified space (e.g., fact  
 211 and sentiment metric space). FSDS network efficiently captures comprehensive fact and sentiment  
 212 sarcasm information for multi-modal sarcasm detection tasks via deep metric learning.  
 213

#### 3.2.1 FACT INCONGRUITY SUB-NETWORK

214 Facts describe the existence of objects or events, which are hidden in semantics. Fact incongruity  
 215 refers to the incongruity between factual semantic information in multi-modal data. FISN aims to  
 216

capture intra-modal and inter-modal incongruities in fact metric spaces as complete multi-modal fact incongruities. Inspired by deep metric learning, we designed the complete multi-modal metric learning (CMLL). CMLL iteratively calculates inter-modal and intra-modal incongruities in the fact metric space via  $N$  CMLL units, which is the designed basic component of CMLL.

Each CMLL unit computes the incongruity of all text-text, text-image, and image-image representation pairs in a unified fact metric space, and updates them via dynamic separation and non-linear adjustment to obtain more discriminative incongruity representations. Thus, it will gradually obtain complete multi-modal incongruities after  $N$  CMLL units. It mainly consists of initial incongruity computation, dynamic separation, and non-linear adjustment.

**Initial incongruity computation** aims to align the unified fact representation space  $\mathcal{S}^F$  into the unified fact incongruity metric space  $(\mathcal{S}^F, \mathcal{I}^F)$ . It computes initial incongruity by measuring the Euclidean distance between each pair of representations. The representation pairs in the unified space  $\mathcal{S}^F$  include text-text, text-image, and image-image. Larger distances indicate higher degrees of incongruity. This establishes a clear reference for reliable updating of the representation of the complete multi-modal incongruity. The incongruity between  $\mathbf{r}_u^F \in \mathcal{S}^F$  and  $\mathbf{r}_v^F \in \mathcal{S}^F$  can be calculated as follows:

$$\mathcal{I}_{u,v}^F = \|\mathbf{r}_u^F - \mathbf{r}_v^F\| \quad (3)$$

**Dynamic separation** enhances the discriminative performance by increasing the separation between incongruous representations. Each representation  $\mathbf{r}_u^F \in \mathcal{S}^F$  receives directional clues from other representation. We employ adaptive weights to reinforce representation itself by relatively congruous representations and prevent assimilation from its incongruous representations. The adaptive weights between each pair of representation  $\mathbf{r}_u^F \in \mathcal{S}^F$  and  $\mathbf{r}_v^F \in \mathcal{S}^F$  in the  $i$ -th CMLL unit are calculated as follows:

$$\mathbf{w}_{u,v}^{F,(i-1)} = \frac{\exp(-(\mathcal{I}_{u,v}^{F,(i-1)})^2)}{\sum_{\mathbf{r}_{v'}^F \in \mathcal{S}^{F,(i-1)}} \exp(-(\mathcal{I}_{u,v'}^{F,(i-1)})^2)} \quad (4)$$

followed by the representation  $\mathbf{r}_u^F \in \mathcal{S}^F$  in the  $i$ -th CMLL unit update:

$$\mathbf{r}_u^{F,(i-1)'} = \mathbf{r}_u^{F,(i-1)} + \sum_{\mathbf{r}_v^F \in \mathcal{S}^{F,(i-1)}} \mathbf{w}_{u,v}^{F,(i-1)} (\mathbf{r}_v^{F,(i-1)} - \mathbf{0}) \quad (5)$$

**Non-linear adjustment** enhances the discriminative performance by refining the topology of the fact metric space. It leverages a simple yet effective deep learning approach to handle complex incongruity relationships and improve the robustness of incongruity discrimination. The representation  $\mathbf{r}_u^F \in \mathcal{S}^F$  in the  $i$ -th CMLL unit update:

$$\mathbf{r}_u^{F,(i)} = FFN(\mathbf{r}_u^{F,(i-1)'}) \quad (6)$$

After the iterative calculating of  $N$  CMLL units, the distance between strongly incongruous representations increases more significantly than that between weakly incongruous ones. FISN obtains more discriminative incongruity representations.

The core pair aggregation module (CPAM) aggregates the fact sarcasm information to output. It firstly screens the target incongruity in the fact incongruity metric space to capture complete multi-modal fact incongruity. The incongruous representation pairs of the fact metric space are summarized into three categories: intra-modal image pairs  $(\mathbf{r}_I^F, \mathbf{r}_I^F)$ , inter-modal pairs  $(\mathbf{r}_T^F, \mathbf{r}_I^F)$ , and intra-modal text pairs  $(\mathbf{r}_T^F, \mathbf{r}_T^F)$ . After selecting the most incongruous pair in each category, we further select the overall most incongruous pair. It helps the CMLL-Net focus on the most significant signal that is most likely to convey sarcasm. The unified metric space captures incongruities in just one efficient step. The capturing in the CPAM is formalized as follows:

$$(\mathbf{r}_{u'}^F, \mathbf{r}_{v'}^F) = \arg \max_{(\mathbf{r}_u^F, \mathbf{r}_v^F) \in \mathcal{S}^F \times \mathcal{S}^F} (\mathcal{I}_{I,I}^F, \mathcal{I}_{T,I}^F, \mathcal{I}_{T,T}^F) \quad (7)$$

CPAM takes the overall most incongruous pair as local information and fuses it with the center of the global representation space to produce an embedding of FISN called fact incongruity (FI).

$$\mathbf{f}_{FI} = (\mathbf{r}_{u'}^F, \mathbf{r}_{v'}^F) \oplus \mathcal{S}_c^F \quad (8)$$

270    3.2.2 SENTIMENT INCONGRUITY SUB-NETWORK  
 271

272    Sarcasm often implicitly expresses dissatisfaction in a positive manner. For instance, the statement  
 273    "What a wonderful day!" alongside an image of a rainy day produces a sentiment incongruity. The  
 274    word "wonderful" expresses a positive sentiment, but the rain in the image shows a negative impres-  
 275    sion. SISN aims to capture intra-modal and inter-modal incongruities in sentiment metric spaces as  
 276    complete multi-modal sentiment incongruities.

277    Similar to FISN, SISN uses CMML to capture complete multi-modal incongruities via  $N$  CMML  
 278    units. Since CMML units in both FISN and SISN perform metric learning in metric spaces, we  
 279    leverage the perspectives of the metric space to distinguish the type of incongruities being captured.

280    Each CMML unit operates within the sentiment metric space  $(\mathcal{S}^S, \mathcal{I}^S)$ . It mainly consists of initial  
 281    incongruity computation, dynamic separation, and non-linear adjustment.

282    In **initial incongruity computation**. The incongruity between  $\mathbf{r}_u^S \in \mathcal{S}^S$  and  $\mathbf{r}_v^S \in \mathcal{S}^S$  can be  
 283    calculated as follows:

$$\mathcal{I}_{u,v}^S = \|\mathbf{r}_u^S - \mathbf{r}_v^S\| \quad (9)$$

284    In **dynamic sepration**, The adaptive weights between each pair of representation  $\mathbf{r}_u^S \in \mathcal{S}^S$  and  
 285     $\mathbf{r}_v^S \in \mathcal{S}^S$  in the  $i$ -th CMML unit are calculated as follows:

$$\mathbf{w}_{u,v}^{S,(i-1)} = \frac{\exp\left(-(\mathcal{I}_{u,v}^{S,(i-1)})^2\right)}{\sum_{\mathbf{r}_{v'}^S \in \mathcal{S}^S} \exp\left(-(\mathcal{I}_{u,v'}^{S,(i-1)})^2\right)} \quad (10)$$

291    followed by the representation  $\mathbf{r}_u^S \in \mathcal{S}^S$  in the  $i$ -th CMML unit update:

$$\mathbf{r}_u^{S,(i-1)'} = \mathbf{r}_u^{S,(i-1)} + \sum_{\mathbf{r}_v^S \in \mathcal{S}^S} \mathbf{w}_{u,v}^{S,(i-1)} (\mathbf{r}_v^{S,(i-1)} - \mathbf{0}) \quad (11)$$

295    In **non-linear adjustment**, the representation  $\mathbf{r}_u^S \in \mathcal{S}^S$  in the  $i$ -th CMML unit update:

$$\mathbf{r}_u^{S,(i)} = FFN\left(\mathbf{r}_u^{S,(i-1)'}\right) \quad (12)$$

298    After the iterative calculating of  $N$  CMML units, SISN obtains more discriminative incongruity  
 299    representations.

300    The core pair aggregation module (CPAM) aggregates the sentiment sarcasm information to out-  
 301    put. The incongruous representation pairs of the sentiment metric space are summarized into three  
 302    categories: intra-modal image pairs  $(\mathbf{r}_I^S, \mathbf{r}_I^S)$ , inter-modal pairs  $(\mathbf{r}_T^S, \mathbf{r}_I^S)$ , and intra-modal text pairs  
 303     $(\mathbf{r}_T^S, \mathbf{r}_T^S)$ . The capturing in the CPAM is formalized as follows:

$$(\mathbf{r}_{u'}^S, \mathbf{r}_{v'}^S) = \arg \max_{(\mathbf{r}_u^S, \mathbf{r}_v^S) \in \mathcal{S}^S \times \mathcal{S}^S} (\mathcal{I}_{I,I}^S, \mathcal{I}_{T,I}^S, \mathcal{I}_{T,T}^S) \quad (13)$$

306    The sentiment incongruity (SI) of SISN is computed as:

$$\mathbf{f}_{SI} = (\mathbf{s}_{u'}^S, \mathbf{s}_{v'}^S) \oplus \mathcal{S}_c^S \quad (14)$$

309    3.3 INCONGRUITY FUSION MODULE  
 310

311    We fuse the FI denoted as  $\mathbf{f}_{FI}$  with the SI denoted as  $\mathbf{f}_{SI}$  as the final incongruity and send it to the  
 312    prediction layer. The total loss function balances the contributions of the fact-related loss, sentiment-  
 313    related loss, and prediction loss:

$$\mathcal{L} = \alpha \mathcal{L}^F + \alpha \mathcal{L}^S + (1 - 2\alpha) \mathcal{L}^{\text{pred}} \quad (15)$$

316    4 EXPERIMENT  
 317

318    4.1 DATASET AND EVALUATION METRICS  
 319

320    We conduct experiments on the publicly available **Multimodal Sarcasm Detection** (MSD) dataset  
 321    (Cai et al., 2019). Each sample in the dataset consists of text-image pairs. Samples expressing sar-  
 322    casm are labeled as positive, and those without sarcasm are labeled as negative. Following previous  
 323    works (Cai et al., 2019; Xu et al., 2020; Liang et al., 2022), we report accuracy, precision, recall,  
 binary-average, and macro-average results for evaluation.

324    **4.2 IMPLEMENTATION DETAILS**

325

326    We used the CLIP ViT-B/32 model (Radford et al., 2021) with frozen parameters for unified token-  
 327    level image and text feature extraction. For the fact stream, we employed YOLO v10-s with frozen  
 328    parameters (Wang et al., 2024a). The CMML-Net was trained using AdamW with a learning rate  
 329    set to 1e-4, weight decay at 1e-4, and  $\alpha$  at 7.5%, over ten epochs.

330    **4.3 BASELINE MODELS**

331

332    To evaluate the performance of CMML-Net, we compare it against several state-of-the-art base-  
 333    lines, categorized into image-modality, text-modality, and multi-modal methods. **Image-modality**  
 334    **methods:** ResNet (Cai et al., 2019), ViT (Dosovitskiy, 2020). **Text-modality methods:** Bi-LSTM  
 335    (Graves & Schmidhuber, 2005), SIARN (Tay et al., 2018), SMSD (Xiong et al., 2019), BERT (Ken-  
 336    ton & Toutanova, 2019). **Multi-modal methods:** HFM (Cai et al., 2019), InCrossMGs (Liang et al.,  
 337    2021), CMGCN (Liang et al., 2022), Att-BERT (Pan et al., 2020), DIP (Wen et al., 2023), Knowl-  
 338    eNet (Yue et al., 2023), FSICN (Lu et al., 2024), Mumu (Wang et al., 2024b), **Multi-view CLIP** (Qin  
 339    et al., 2023), **DMSD-CL** (Jia et al., 2024), **G2SAM** (Wei et al., 2024). The details of these methods  
 340    have been described in the related work section.

341

342    **4.4 MAIN RESULT**

343

344    Table 1 presents the performance comparison of our proposed method against other state-of-the-  
 345    art (SOTA) approaches. CMML-Net achieves the best performance across all evaluation metrics,  
 346    demonstrating its ability to more comprehensively and accurately capture sarcasm information.  
 347    Specifically, CMML-Net outperforms unimodal approaches due to the complementary nature of  
 348    multi-modal data. Among unimodal methods, text-based approaches perform better than image-  
 349    based ones, largely because of the abstract nature and implicit details in images. Therefore, CMML-  
 350    Net employs a Yolo-task representation extraction to learn fine-grained factual semantics from im-  
 351    ages.

352    CMML-Net outperforms existing SOTA method MuMu (Wang et al., 2024b) by a significant mar-  
 353    gin—achieving +1.4% (90.73% vs. 92.04%), +1.8% (88.62% vs. 90.25%), and +1.5% (90.40%  
 354    vs. 91.76%) on accuracy, binary-F1, and macro-F1 scores, respectively. Although MuMu utilizes  
 355    the same feature extractor as ours, MuMu is limited because of the incomplete multi-modal incon-  
 356    gruities. The performance gap between CMML-Net and these models highlights the effectiveness  
 357    of CMML-Net in capturing complete multi-modal incongruities in fact and sentiment perspectives.

358    **4.5 ABLATION STUDY**

359

360    We conducted an ablation study to assess the im-  
 361    pact of each module, as shown in Table 2. First,  
 362    the absence of FISN and SISN also led to per-  
 363    formance degradation. It demonstrates the bi-  
 364    ased performance of capturing multi-modal in-  
 365    congruity only from a fact perspective or senti-  
 366    ment perspective.

367    Additionally, Removing the Yolo-task fact repre-  
 368    sentation extraction led to a drop in accuracy and  
 369    macro-F1, underscoring the importance of learn-  
 370    ing fine-grained factual semantics via Yolo-task.  
 371    Yolo-task improves the quality of fact repres-  
 372    entation space and prevents the performance from decreasing because of the collapse between fact  
 373    representation and sentiment representation. In high-quality and separate fact and sentiment metric  
 374    space, CMML-Net achieves more significant performance gains by deep metric learning.

375    Ablating CMML units showed further performance reductions. CMML units play a critical role in  
 376    calculating complete multi-modal incongruity in fact and sentiment metric space. The model with  
 377    CMML units obtains powerful discriminative performance and handles the biased performance by  
 capturing complete multi-modal incongruity in fact and sentiment perspectives.

Table 2: Ablation study results.

Model Name	ACC(%)	F1(%)
w/o FISN	91.23	90.87
w/o SISN	91.28	90.89
w/o Yolo-task	91.19	90.85
w/o CMML units	90.98	90.68
Full Model	<b>92.04</b>	<b>91.76</b>

Modality	Model	ACC(%)	Binary-Average			Macro-Average		
			P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Image	Resnet	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT	67.83	57.93	70.07	63.43	65.69	71.35	68.40
Text	Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	SIARN	80.57	75.55	75.70	75.63	80.34	78.81	79.57
	SMSD	80.90	76.46	75.18	75.82	80.87	78.20	79.51
	BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
Image+Text	HFM	86.63	83.84	84.18	84.01	86.24	86.28	86.26
	InCrossMGs	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	CMGCN	87.55	83.63	84.69	84.16	87.02	86.97	87.00
	Att-Bert	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	DIP	89.59	87.76	86.58	87.17	88.46	89.13	89.01
	KnowleNet	88.87	88.59	84.18	86.33	88.83	88.21	88.51
	<b>Multi-view CLIP</b>	<b>88.33</b>	<b>82.66</b>	<b>88.65</b>	<b>85.55</b>	-	-	-
	DMSD-CL	<b>88.95</b>	<b>84.89</b>	<b>87.90</b>	<b>86.37</b>	<b>88.35</b>	<b>88.77</b>	<b>88.54</b>
	G2SAM	<b>90.48</b>	<b>87.95</b>	<b>89.02</b>	<b>88.48</b>	<b>89.44</b>	<b>89.79</b>	<b>89.65</b>
	FSICN	90.55	<b>89.93</b>	<b>89.51</b>	<b>89.72</b>	90.16	<b>90.42</b>	90.29
	MuMu	90.73	88.81	88.44	88.62	90.43	90.37	90.40
	<b>Ours*</b>	<b>92.04*</b>	<b>90.21*</b>	<b>90.30*</b>	<b>90.25*</b>	<b>91.75*</b>	<b>91.77*</b>	<b>91.76*</b>

Table 1: Results with \* denote the significance tests of our CMML-Net over the baseline models at p-value < 0.01. The best results are highlighted in boldface, while the second-best results are underlined.

#### 4.6 IMPACT OF YOLO-TASK REPRESENTATION EXTRACTION ON REPRESENTATION SPACE

We conducted an experiment to evaluate how Yolo-task representation extraction influences the separation of initial fact and sentiment representation spaces. To our knowledge, none of the existing fact-sentiment dual-stream networks research explicitly guides representations of fact sub-network in multi-modal sarcasm detection (Wen et al., 2023; Lu et al., 2024). We applied UMAP for dimensional reduction while preserving topology (Figure 3). The left figure shows the fact-sentiment representation space without the guidance information produced from the Yolo-task, and the right figure shows the fact-sentiment representation space after introducing Yolo-task representation extraction. By analyzing the metric spaces, we observe that the lack of guidance information produced from the Yolo-task causes the FISN to collapse into the SISN. This limits the FISN’s ability to capture incongruity at the fine-grained factual semantics level. This also weakens the SISN’s contribution and leads to overall performance degradation. The Yolo-task representation learning module prevents this collapse and enhances the model’s ability to maintain distinct representation spaces.

In summary, Yolo-task representation extraction significantly improves the separation between fact and sentiment representation spaces. We further validated this by training GAN models to distinguish between the two spaces. The model with Yolo-task representation extraction achieved a much higher accuracy of 93.17%, compared to 66.52% without Yolo-task representation extraction, confirming a 26.59% improvement in performance. This further demonstrates the effectiveness of Yolo-task representation extraction in preserving the integrity of FISN and SISN.

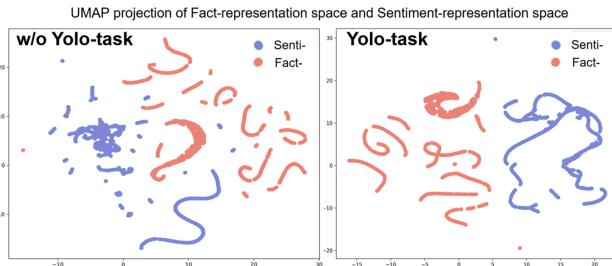


Figure 3: The fact-sentiment representation space. Model without Yolo-task (left) vs. with Yolo-task (right).

## 432 4.7 ROLE OF CMML UNITS IN CAPTURING COMPLETE MULTI-MODAL INCONGRUITY

433  
 434 We designed an experiment to verify  
 435 the role of the CMML units in  
 436 capturing complete multi-modal in-  
 437 congruities by tracking the per-  
 438 formance of the full model and the  
 439 model without the CMML units dur-  
 440 ing training (shown in Figure 4). The  
 441 model without CMML units lacks  
 442 the mechanism to iteratively calculate  
 443 inter-modal and intra-modal incon-  
 444 gruities in the fact and sentiment met-  
 445 ric space. It captures incongruities  
 446 with the initial discriminative per-  
 447 formance. The stacked bar chart dis-  
 448 plays the proportions of sentiment  
 449 and fact incongruities, which are fur-  
 450 ther divided into inter-modal, intra-  
 451 image, and intra-text incongruities.

452 The data shows that both models ex-  
 453 hibit similar performance trends in  
 454 the early training stages. It indicates  
 455 that initial optimization depends on  
 456 global information from metric space.  
 457 However, as training progresses and the  
 458 models reach the first performance peak,  
 459 the full model enters a refinement phase  
 460 driven by CMML units. During this  
 461 phase,  $N$  CMML units iteratively calcu-  
 462 late inter-modal and intra-modal incon-  
 463 gruities in the fact and sentiment met-  
 464 ric space. The model obtains more dis-  
 465 criminative incongruity representations  
 466 and efficiently captures complete multi-  
 467 modal incongruity. The proportion of intra-image incongruities  
 468 rises and eventually matches that of intra-text incongruities. In contrast,  
 469 the model without CMML units stagnates,  
 470 with its performance fluctuating slightly.  
 471 This demonstrates that CMML units are  
 472 crucial for handling biased performance  
 473 through comprehensive sarcasm infor-  
 474 mation, especially in later training  
 475 stages.

## 476 4.8 MODEL EFFICIENCY ANALYSIS

477 We conducted experiments to validate the efficiency of CMML in capturing complete multi-modal  
 478 incongruities in a unified metric space. We also compare it with the prior methods that have potential ability to capture complete multi-modal incongruities through reconstructing the framework.  
 479 DIP (Wen et al., 2023) identifies inter-modal incongruities using Gaussian distribution differences, but its mechanism cannot be directly applied to intra-modal incongruities. Due to some practical limitations, it is difficult for us to reproduce FSICN (Lu et al., 2024). Att-Bert (Pan et al., 2020) employed separate attention mechanisms, and InCrossMGs (Liang et al., 2021) used distinct GCNs to extract inter-modal and intra-modal incongruities separately. This leads to significant computational overhead.

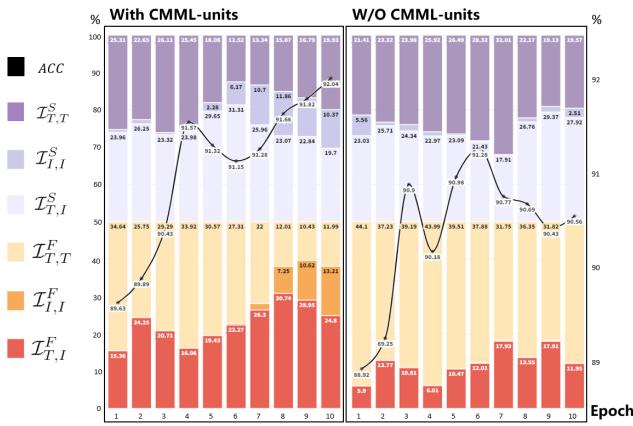


Figure 4: The proportion of max incongruity representation from inter-modal or intra-modal perspectives in both fact and sentiment perspectives over different epochs. Model with CMML units (Left) vs. without CMML units (Right).

Table 3: Model Efficiency Comparision.

Model	Performance		Backend Model	
	ACC(%)	F1(%)	Params	GFLOPS
Att-Dual	90.98	90.65	105.7M ( $\uparrow$ 511%)	204.4 ( $\uparrow$ 527%)
GCN-Dual	90.39	90.06	49.4M ( $\uparrow$ 186%)	127.3 ( $\uparrow$ 290%)
Ours	<b>92.04</b>	<b>91.76</b>	<b>17.3M</b>	<b>32.6</b>

486 For a fair comparison, we replicated previous models based on GCN and attention mechanisms  
 487 (Pan et al., 2020; Liang et al., 2021) within our dual-stream framework and excluded the common  
 488 frontend architecture (CLIP and Yolo) to focus on the backend.

489 As shown in Table 3, CMML-Net demonstrates a significant reduction in computational complexity,  
 490 using 65% fewer parameters and 74% fewer FLOPS compared to the GCN-dual-stream model and  
 491 83% fewer parameters and 84% fewer FLOPS compared to the Attention-dual-stream model. These  
 492 results confirm CMML’s ability to maintain high performance with much lower resource demands.  
 493

## 494 5 CONCLUSION

495 In this paper, we propose CMML-Net, the first work in multi-modal sarcasm detection to introduce  
 496 deep metric learning to explicitly and efficiently capture complete multi-modal incongruities in fact  
 497 and sentiment perspectives. In the fact-sentiment multi-task representation learning module, we  
 498 use Yolo-task and SenticNet-task representation extraction to produce refined fact and sentiment  
 499 text-image representation pairs, respectively. It enhances the model’s ability to maintain distinct  
 500 representation spaces. In the fact-sentiment dual-stream network consisting of FISN and SISN,  
 501 we designed a CMML to iteratively calculate inter-modal and intra-modal incongruities in fact  
 502 and sentiment perspectives a unified space (e.g., fact and sentiment metric space) by a complete  
 503 multi-modal metric learning. It then explicitly and efficiently captures complete multi-modal incon-  
 504 gruities. CMML-Net handles the biased performance of sarcasm detection models by comprehen-  
 505 sive sarcasm information. The state-of-the-art performance on the widely-used dataset demonstrates  
 506 CMML-Net’s superiority.

## 507 6 REPRODUCIBILITY STATEMENT

511 We provide a detailed description of the implementation in Section 4.2. Additionally, we have open-  
 512 sourced our codebase to facilitate the reproduction of our results, which are available on the project  
 513 website: [<https://anonymous.4open.science/r/CMML-Net-873E>].

## 514 REFERENCES

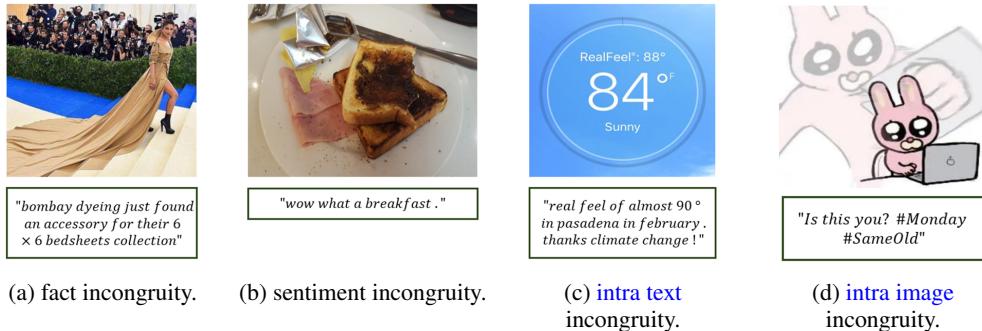
- 517 Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical  
 518 fusion model. In *Proceedings of the 57th annual meeting of the association for computational  
 519 linguistics*, pp. 2506–2515, 2019.
- 520 Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate  
 521 in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial  
 522 Intelligence*, volume 36, pp. 10563–10571, 2022.
- 523 Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale.  
 524 *arXiv preprint arXiv:2010.11929*, 2020.
- 525 Raymond W Gibbs. On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 115(1):3, 1986.
- 528 Amir Globerson and Sam Roweis. Metric learning by collapsing classes. *Advances in neural  
 529 information processing systems*, 18, 2005.
- 531 Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter:  
 532 a closer look. In *Proceedings of the 49th annual meeting of the association for computational  
 533 linguistics: human language technologies*, pp. 581–586, 2011.
- 534 Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm  
 535 and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.
- 536 Herbert Paul Grice. Further notes on logic and conversation. *Syntax and semantics*, 9, 1978.
- 538 Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant  
 539 mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition  
 (CVPR’06)*, volume 2, pp. 1735–1742. IEEE, 2006.

- 540 Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Similarity-based pattern  
 541 recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October  
 542 12-14, 2015. Proceedings 3, pp. 84–92. Springer, 2015.
- 543
- 544 Mengzhao Jia, Can Xie, and Liqiang Jing. Debiasing multimodal sarcasm detection with contrastive  
 545 learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp.  
 546 18354–18362, 2024.
- 547
- 548 Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. Harnessing context incongruity for sar-  
 549 casm detection. In Proceedings of the 53rd Annual Meeting of the Association for Computational  
Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume  
 550 2: Short Papers), pp. 757–762, 2015.
- 551
- 552 Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep  
 553 bidirectional transformers for language understanding. In Proceedings of naacl-HLT, volume 1,  
 554 pp. 2, 2019.
- 555
- 556 Y Alex Kolchinski and Christopher Potts. Representing social media users for sarcasm detection.  
 557 arXiv preprint arXiv:1808.08470, 2018.
- 558
- 559 Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Rui Feng Xu. Multi-modal sarcasm  
 560 detection with interactive in-modal and cross-modal graphs. In Proceedings of the 29th ACM  
international conference on multimedia, pp. 4707–4715, 2021.
- 561
- 562 Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Rui Feng Xu.  
 563 Multi-modal sarcasm detection via cross-modal graph convolutional network. In Proceedings  
564 of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long  
Papers), volume 1, pp. 1767–1777. Association for Computational Linguistics, 2022.
- 565
- 566 Qiang Lu, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. Fact-sentiment incongruity combination  
 567 network for multimodal sarcasm detection. Information Fusion, 104:102203, 2024.
- 568
- 569 Krishnan Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. A multitask framework for  
 570 sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed  
 571 memes. In Proceedings of the 45th International ACM SIGIR Conference on Research and  
572 Development in Information Retrieval, pp. 1739–1749, 2022.
- 573
- 574 Skye McDonald. Exploring the process of inference generation in sarcasm: A review of normal and  
 575 clinical studies. Brain and language, 68(3):486–506, 1999.
- 576
- 577 Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. Modeling intra and inter-  
 578 modality incongruity for multi-modal sarcasm detection. In Findings of the Association for  
579 Computational Linguistics: EMNLP 2020, pp. 1383–1392, 2020.
- 580
- 581 Liwen Peng, Songlei Jian, Dongsheng Li, and Siqi Shen. Mrml: Multimodal rumor detection by  
 582 deep metric learning. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech  
583 and Signal Processing (ICASSP), pp. 1–5. IEEE, 2023.
- 584
- 585 Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and  
 586 Rui Feng Xu. Mmsd2. 0: towards a reliable multi-modal sarcasm detection system. arXiv preprint  
587 arXiv:2307.07135, 2023.
- 588
- 589 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
 590 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual  
 591 models from natural language supervision. In International conference on machine learning, pp.  
 592 8748–8763. PMLR, 2021.
- 593
- 594 Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang.  
 595 Sarcasm as contrast between a positive sentiment and negative situation. In Proceedings of the  
596 2013 conference on empirical methods in natural language processing, pp. 704–714, 2013.

- 594 Chhavi Sharma, William Pak, Scott, Deepesh Bhageria, Amitava Das, Soujanya Poria, Tan-  
 595 moy Chakraborty, and Björn Gambäck. Task report: Memotion analysis 1.0 @semeval 2020:  
 596 The visuo-lingual metaphor! In Proceedings of the 14th International Workshop on Semantic  
 597 Evaluation (SemEval-2020), Barcelona, Spain, Sep 2020. Association for Computational Lin-  
 598 guistics.
- 599 Dan Sperber and Deirdre Wilson. Précis of relevance: Communication and cognition. Behavioral  
 600 and brain sciences, 10(4):697–710, 1987.
- 601 Juan Luis Suárez, Salvador García, and Francisco Herrera. A tutorial on distance metric learn-  
 602 ing: Mathematical foundations, algorithms, experimental analysis, prospects and challenges.  
 603 Neurocomputing, 425:300–322, 2021.
- 604 Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. Reasoning with sarcasm by reading in-  
 605 between. arXiv preprint arXiv:1805.02856, 2018.
- 606 Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10:  
 607 Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458, 2024a.
- 608 Fei Wang and Jimeng Sun. Survey on distance metric learning and dimensionality reduction in data  
 609 mining. Data mining and knowledge discovery, 29(2):534–564, 2015.
- 610 Haosen Wang, Pan Tang, Hanyue Kong, Yilun Jin, Chunqi Wu, and Linghong Zhou. Dhcf: Dual  
 611 disentangled-view hierarchical contrastive learning for fake news detection on social media.  
 612 Information Sciences, 645:119323, 2023.
- 613 Jie Wang, Yan Yang, Yongquan Jiang, Minbo Ma, Zhuyang Xie, and Tianrui Li. Cross-modal  
 614 incongruity aligning and collaborating for multi-modal sarcasm detection. Information Fusion,  
 615 103:102132, 2024b.
- 616 Yiwei Wei, Shaozu Yuan, Hengyang Zhou, Longbiao Wang, Zhiling Yan, Ruosong Yang, and Meng  
 617 Chen. G<sup>^</sup> 2sam: Graph-based global semantic awareness method for multimodal sarcasm detec-  
 618 tion. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 9151–  
 619 9159, 2024.
- 620 Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin  
 621 nearest neighbor classification. Advances in neural information processing systems, 18, 2005.
- 622 Changsong Wen, Guoli Jia, and Jufeng Yang. Dip: Dual incongruity perceiving network for sar-  
 623 casm detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern  
 624 Recognition, pp. 2540–2550, 2023.
- 625 Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. Sarcasm detection with self-matching  
 626 networks and low-rank bilinear pooling. In The world wide web conference, pp. 2115–2124,  
 627 2019.
- 628 Nan Xu, Zhixiong Zeng, and Wenji Mao. Reasoning with multimodal sarcastic tweets via modeling  
 629 cross-modality contrast and semantic association. In Proceedings of the 58th annual meeting of  
 630 the association for computational linguistics, pp. 3777–3786, 2020.
- 631 Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. Knowlenet: Knowledge fusion  
 632 network for multimodal sarcasm detection. Information Fusion, 100:101921, 2023.
- 633 Meishan Zhang, Yue Zhang, and Guohong Fu. Tweet sarcasm detection using deep neural net-  
 634 work. In Proceedings of COLING 2016, the 26th International Conference on Computational  
 635 Linguistics: technical papers, pp. 2449–2460, 2016.
- 636
- 637
- 638
- 639
- 640
- 641
- 642
- 643
- 644
- 645
- 646
- 647

## 648 A EXAMPLES OF THE INCONGRUITY IN SARCASTIC REMARKS 649

650 We demonstrate the effectiveness of capturing complete multi-modal incongruity in fact and senti-  
 651 ment perspectives through some specific sarcastic remarks (Figure 5). Incomplete incongruity will  
 652 lead to biased performance due to the lack of sarcastic information.  
 653



654  
 655  
 656  
 657  
 658  
 659  
 660 Figure 5: Examples of the incongruity in sarcastic remarks.  
 661  
 662  
 663  
 664

### 665 [Sample a in Figure 5] (Fact Incongruity) 666

667 **Reason:** In the text, the image of a woman wearing a floor-length dress is described as “Bombay  
 668 Dyeing just found an accessory for their 6×6 bedsheets collection.” Although the text does not ex-  
 669 plicitly mention the dress, it metaphorically compares it to a bedsheet. This implicit fact incongruity  
 670 satirizes the impracticality and excessive extravagance of the woman’s attire.  
 671

### 672 [Sample b in Figure 5] (Sentiment Incongruity)

673 **Reason:** The image shows a burnt piece of bread, which naturally evokes dissatisfaction. In contrast,  
 674 the text enthusiastically states, “Wow, what a breakfast.” This sentiment incongruity highlights the  
 675 sarcasm of the stark contrast between the quality of the breakfast and the exaggerated expectation.  
 676

### 677 [Sample c in Figure 5] (Intra-modal Text Incongruity)

678 **Reason:** The image displays a weather forecast with “RealFeel: 88°, Now: 84°, Sunny,” which is  
 679 congruous with the text’s statement, “real feel of almost 90° in Pasadena in February.” However, the  
 680 incongruity arises within the text itself: the first sentence conveys dissatisfaction with the unusually  
 681 high temperature in February, while the second sarcastically states, “thanks climate change.” This  
 682 sentimental reversal highlights the sarcasm of the global warming phenomenon.  
 683

### 684 [Sample d in Figure 5] (Intra-modal Image Incongruity)

685 **Reason:** The text “#Monday #SameOld” provides contextual background for the image, while “Is  
 686 this you?” encourages viewers to reflect on the scene. The text complements the image, establishing  
 687 congruity between the modalities. However, the incongruity within the image arises from the anthro-  
 688 pomorphic rabbit appearing to work diligently on a computer while internally imagining smashing  
 689 it. This contrast satirizes the conflict in modern work environments between outward composure  
 690 and suppressed frustration.  
 691  
 692  
 693  
 694  
 695  
 696  
 697  
 698  
 699  
 700  
 701

702      **B EXPERIMENTS RESULTS ON MMSD2.0 AND DMSD**  
 703

704      To verify the generalization of CMML-Net, we conducted further experiments on the MMSD2.0  
 705      and DMSD datasets and report the main results below:  
 706

Modality	Model	ACC(%)	P(%)	R(%)	F1(%)
Text-Only	TextCNN	71.61	64.62	75.22	69.52
	BiLSTM	72.48	68.02	68.08	68.05
	SMSD	73.56	68.45	71.55	69.97
Image-Only	ResNet	65.50	61.17	54.39	57.58
	ViT	72.02	65.26	74.83	69.72
Multimodal	HFM	70.57	64.84	69.05	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	72.07	72.25
	DynRT-Net	71.40	71.80	72.17	71.34
	Multi-view CLIP (Frozen)	84.72	-	-	83.64
	Multi-view CLIP (Full Finetuned)	85.64	80.33	88.24	84.10
	<b>Ours</b>	<b>85.83*</b>	<b>80.58*</b>	<b>88.30*</b>	<b>84.26*</b>

722      Table 4: Model performance on MMSD2.0 dataset. Results with \* indicate statistical significance  
 723      over the baseline models at p-value < 0.01. The best results are highlighted in boldface, while the  
 724      second-best results are underlined.  
 725

Modality	Model	ACC(%)	P(%)	R(%)	F1(%)
Text-Only	TextCNN	37.25	37.30	36.71	36.58
	BiLSTM	34.50	33.20	32.77	32.94
	BERT	21.25	22.22	22.28	21.25
	RoBERTa	29.50	128.07	27.34	27.64
Image-Only	ResNet	28.25	27.87	27.04	27.36
	ViT	22.00	22.53	21.36	21.55
Multimodal	Res-BERT	20.75	21.62	20.77	20.60
	Att-Bert	28.25	27.50	26.46	26.69
	HKE	37.50	37.90	37.36	37.04
	CMGCN	34.25	35.52	35.22	34.20
	DMSD-CL	70.25	70.41	71.34	69.96
	<b>Ours</b>	<b>75.75*</b>	<b>76.81*</b>	<b>71.72*</b>	<b>72.53*</b>

741      Table 5: Model performance on DMSD dataset. Results with \* indicate statistical significance over  
 742      the baseline models at p-value < 0.05. The best results are highlighted in boldface, while the second-  
 743      best results are underlined.  
 744

745      Due to time constraints, we used the frozen version of CLIP as the encoder. Our CMML-Net still  
 746      outperforms the fully fine-tuned Multi-view CLIP on MMSD2.0 (Table 4). CMML-Net achieves  
 747      state-of-the-art (SOTA) results on the DMSD dataset (Table 5).  
 748

749      Based on the results from both the MMSD2.0 and DMSD datasets (Table 4, Table 5), CMML-  
 750      Net consistently demonstrates its effectiveness in multimodal sarcasm detection. By explicitly and  
 751      efficiently capturing complete multi-modal incongruities in both fact and sentiment perspectives, our  
 752      approach achieves competitive performance without requiring extensive fine-tuning or large-scale  
 753      architectures. This highlights CMML-Net’s robustness and strong potential for broader applications  
 754      in multimodal tasks.

756      **C EXPERIMENTAL RESULTS OF MODEL WITH RoBERTa ENCODER ON MSD**  
 757      **DATASET**

759      We provide results with RoBERTa as the text encoder, shown in Table 6.

761 <b>Model</b>	ACC(%)	762 <b>Binary-Average</b>			763 <b>Macro-Average</b>		
		764 <b>P(%)</b>	765 <b>R(%)</b>	766 <b>F1(%)</b>	767 <b>P(%)</b>	768 <b>R(%)</b>	769 <b>F1(%)</b>
MILNet	89.50	85.16	89.16	87.11	88.88	89.44	89.12
DynRT-Net	93.59	93.06	93.60	93.31	-	-	-
FSICN+RoBERTa	94.71	93.62	93.28	93.45	-	-	-
<b>Ours+RoBERTa</b>	<b>97.05*</b>	<b>99.45*</b>	<b>93.29</b>	<b>96.27*</b>	<b>97.51*</b>	<b>96.47*</b>	<b>96.92*</b>

770      Table 6: Supplemented results on MSD dataset with RoBERTa encoder. Results with \* indicate  
 771      statistical significance over the baseline models at p-value < 0.01. The best results are highlighted  
 772      in boldface, while the second-best results are underlined.

773      By supplementing these baselines and demonstrating consistently state-of-the-art (SOTA) results  
 774      (Table 6), our experimental analysis substantiates the superiority of CMML-Net in multimodal sar-  
 775      casm detection tasks.

776  
 777  
 778  
 779  
 780  
 781  
 782  
 783  
 784  
 785  
 786  
 787  
 788  
 789  
 790  
 791  
 792  
 793  
 794  
 795  
 796  
 797  
 798  
 799  
 800  
 801  
 802  
 803  
 804  
 805  
 806  
 807  
 808  
 809

810  
811     **D DISCUSSION OF POTENTIAL APPLICATIONS AND BROADER**  
812     **IMPLICATIONS**

813     Our proposed complete multi-modal metric learning method can jointly and explicitly calculate  
814     inter-modal and intra-modal incongruity. It is applicable to multimodal tasks such as fake news  
815     detection and sentiment transition analysis. It can jointly and efficiently reveal the key incongruous  
816     features in these tasks from both inter-modal and intra-modal aspects, thereby improving the perfor-  
817     mance of the model. This study can provide an important reference for researchers in these similar  
818     tasks and promote the further development of multimodal learning methods.

819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

---

**E ADDITIONAL RELATED WORK: METRIC LEARNING**

Metric learning is a technique for measuring the similarity between objects based on distance metrics. Traditional methods usually transform the original feature space into a representation where the distance can capture meaningful relationships. These methods usually include Mahalanobis distance (Globerson & Roweis, 2005; Wang & Sun, 2015; Weinberger et al., 2005) and rely on linear transformations (e.g., symmetric positive definite matrices) to project the data into Euclidean space. They are limited by their reliance on predefined distance functions.

Deep metric learning extends this concept by leveraging nonlinear transformations through deep neural networks. This approach creates a flexible embedding space. It enables the model to minimize the distance between similar samples and maximize the separation between different samples (Peng et al., 2023; Wang et al., 2023). Deep metric learning has been successful in various applications, including image text retrieval, text classification, face recognition, and multimodal data representation (Suárez et al., 2021). Existing works widely adopt architectures such as Siamese networks and loss functions (e.g., triplet loss (Hoffer & Ailon, 2015) and contrastive loss (Hadsell et al., 2006)) to effectively capture the relationship between pairs or groups of samples.

We introduce the concepts of deep metric learning to calculate incongruity at the representation level rather than at the sample level. Our method iteratively calculates complete multi-modal incongruity to capture the subtle relationships between representations for multi-modal sarcasm detection. This approach generalizes deep metric learning concepts to capture more complex relationships, offering a broader and more adaptable solution for multimodal learning tasks

864  
865866 Metric learning is a technique for measuring the similarity between objects based on distance met-  
867 rics. Traditional methods usually transform the original feature space into a representation where  
868 the distance can capture meaningful relationships. These methods usually include Mahalanobis dis-  
869 tance (Globerson & Roweis, 2005; Wang & Sun, 2015; Weinberger et al., 2005) and rely on linear  
870 transformations (e.g., symmetric positive definite matrices) to project the data into Euclidean space.  
871 They are limited by their reliance on predefined distance functions.872 Deep metric learning extends this concept by leveraging nonlinear transformations through deep  
873 neural networks. This approach creates a flexible embedding space. It enables the model to mini-  
874 mize the distance between similar samples and maximize the separation between different samples  
875 (Peng et al., 2023; Wang et al., 2023). Deep metric learning has been successful in various ap-  
876 plications, including image text retrieval, text classification, face recognition, and multimodal data  
877 representation (Suárez et al., 2021). Existing works widely adopt architectures such as Siamese  
878 networks and loss functions (e.g., triplet loss (Hoffer & Ailon, 2015) and contrastive loss (Hadsell  
879 et al., 2006)) to effectively capture the relationship between pairs or groups of samples.880 We introduce the concepts of deep metric learning to calculate incongruity at the representation level  
881 rather than at the sample level. Our method iteratively calculates complete multi-modal incongruity  
882 to capture the subtle relationships between representations for multi-modal sarcasm detection. This  
883 approach generalizes deep metric learning concepts to capture more complex relationships, offering  
884 a broader and more adaptable solution for multimodal learning tasks

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918  
 919   **F EXPERIMENTAL RESULTS OF MODEL WITH DIFFERENT BACKBONES ON**  
 920   **MSD DATASET**

921   To investigate the impact of different backbones of Yolo-task on the MSD, we conducted experi-  
 922   ments using various versions of YOLOv10.

924   Table 7: Results on MSD Dataset with Different Backbones of Yolo-task  
 925

Backbone	ACC (%)	F1 (%)	#Params	FLOPs
w/o YOLO-task	91.19	90.85	-	-
YOLO v10-N	91.61	91.27	2.3M	6.7G
YOLO v10-M	91.61	91.31	15.4M	59.1G
YOLO v10-B	91.53	91.23	19.1M	92.0G
YOLO v10-L	91.53	91.24	24.4M	120.3G
YOLO v10-X	91.83	91.52	29.5M	160.4G
<b>YOLO v10-S (Ours)</b>	<b>92.04</b>	<b>91.76</b>	<b>7.2M</b>	<b>21.6G</b>

935   The experimental results suggest that the performance of the framework is primarily influenced by  
 936   the guiding role of the YOLO-task, while the complexity of the chosen object detection backbone  
 937   plays a less significant role.

938  
 939  
 940  
 941  
 942  
 943  
 944  
 945  
 946  
 947  
 948  
 949  
 950  
 951  
 952  
 953  
 954  
 955  
 956  
 957  
 958  
 959  
 960  
 961  
 962  
 963  
 964  
 965  
 966  
 967  
 968  
 969  
 970  
 971