DGLF: A Dual Graph-based Learning Framework for Multi-modal Sarcasm Detection

Zhihong Zhu[1,†], **Kefan Shen**[1,†], **Zhaorun Chen**[3], **Yunyan Zhang**[2], **Yuyan Chen**[4]
Xiaoqi Jiao[5], **Zhongwei Wan**[6], **Shaorong Xie**[1], **Wei Liu**[1,*], **Xian Wu**[2,*], **Yefeng Zheng**[2,7]
[1]School of CES, Shanghai University, [2]Jarvis Research Center, Tencent YouTu Lab
[3]The University of Chicago, [4]Fudan University
[5]Huazhong University of Science and Technology
[6]The Ohio State University, [7]Westlake University
{mrzhuzh, kefanshen}@shu.edu.cn
liuw@shu.edu.cn, kevinxwu@tencent.com

Abstract

Capturing inter-modal incongruities within the text-image pair is a critical challenge in multi-modal sarcasm detection (MSD). Fortunately, graph neural networks (GNNs) have made promising advancements in MSD, which show advantages in explicitly capturing data relationships. Nevertheless, current GNN-based MSD methods do not effectively address some of the inherent limitations of GNNs, which include: 1) neglecting high-order relationships, and 2) underestimating high-frequency messages. In this paper, we propose a **D**ual **G**raph-based **L**earning **F**ramework (DGLF) to address the above two issues. Specifically, we construct a hypergraph to perform high-order aware propagation and a vanilla graph to perform high-frequency enhanced propagation, respectively. We empower GNNs to 1) better capture the inherent and complicated relationships based on the hypergraph and 2) deliver sufficient modeling through high-frequency enhanced messages on the vanilla graph. Besides, we introduce multi-modal fusion information bottleneck to effectively fuse the two learned graph features. Experimental results on two benchmark datasets show that the proposed model outperforms previous state-of-the-art methods.

1 Introduction

Due to the rise of social media platforms such as X and Facebook, multi-modal sarcasm detection (MSD) has garnered increasing research attention. MSD aims to recognize the sarcastic sentiment in multi-modal social posts (Cai et al., 2019), which typically refer to textual sentences accompanying images. Unlike traditional text-only sarcasm detection (Riloff et al., 2013; Poria et al., 2016; Zhang et al., 2016) focusing on inconsistencies in expression within the text, the key objective of MSD is to effectively identify subtle inter-modal inconsistencies in the expression of sentiment within an image-text pair, as shown in Figure 1.

(a) what a gorgeous day

(b) feeding my abs nothing but the best quality beef

Figure 1: Two examples of multi-modal sarcasm. In example (a), the text refers to a "*gorgeous*" day, but the accompanying image shows heavy rain, indicating sarcastic. In example (b), the text mentions "*the best quality beef*", but the image displays fast food beef burgers, suggesting sarcastic as well.

Towards this goal, a group of MSD works attempt to concatenate the textual and visual features to encapsulate sarcastic information (Schifanella et al., 2016), or leverage attention mechanism (Vaswani et al., 2017) to *implicitly* fuse features across modalities based on external knowledge (Cai et al., 2019; Pan et al., 2020). More recently, Graph Neural Networks (GNNs) have achieved remarkable advancements in MSD, showcasing their exceptional ability by *explicitly* extracting structural information (Liang et al., 2021, 2022; Liu et al., 2022). As shown in Figure 2(a), the conventional approach in this paradigm constructs a heterogeneous graph, where each token from both modalities is treated as a node, with similarity-based edge construction or carefully adjusted edge weighting strategies. On this basis, it enables simultaneous modeling of inter- and intra-modal token dependencies through message passing, facilitating tighter entanglement and richer interactions.

Despite the promising progress these GNN-based MSD models have achieved, we discover

[†]Equal contribution.
[*]Corresponding authors: Xian Wu and Wei Liu.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2900–2912
November 12-16, 2024 ©2024 Association for Computational Linguistics

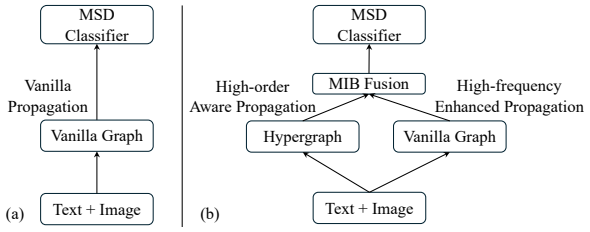

Figure 2: Conceptual comparison of the state-of-the-art methods (a) and the proposed model DGLF (b). "MIB" is short for multi-modal information bottleneck.

that they still suffer from two main issues:

(1) **Neglect of high-order relationship exploration**. They predominantly conceptualize relationships among tokens in pairwise formulations, providing merely an approximation of higher-order relationships through the aggregation of multiple pairs (Feng et al., 2019; Bai et al., 2021). As depicted in Figure 1(b), crucial visual information "*burger*", which aligns with sarcastic textual cues "*best quality beef*", may be scattered across the image (two burgers within the image). While it is feasible to construct edges simultaneously among two visual tokens involving "*burger*" and textual tokens of "*best quality beef*", simplifying these high-order relationships into pairwise formulation might compromise their expressiveness (Sun et al., 2021). Therefore, the sophisticated and nuanced high-order relationships may not be fully captured by existing GNN-based MSD methods.

(2) **Overlooking high-frequency messages exploitation**. The propagation rule of GNNs, characterized by the aggregation and smoothing of messages from neighboring nodes, is widely regarded as an analogy to a fixed low-pass filter (Wu et al., 2019). It predominantly facilitates the flow of low-frequency messages in the graph while significantly attenuating high-frequency messages (Bo et al., 2021). Conversely, in GNNs for MSD, the high-frequency messages may be more vital which reflects discrepancies and inconsistencies in the expression of sentiment. As such, the potential of high-frequency information remains largely unexploited in existing GNN-based MSD frameworks.

To tackle the above issues, as shown in Figure 2(b), we propose a **D**ual **G**raph-based **L**earning **F**ramework termed DGLF for MSD. **For the first issue**, we construct a hypergraph (Feng et al., 2019) with edge-dependent node weights (Chitra and Raphael, 2019) to facilitate high-order aware propagation, where each token from both modalities is represented as a node. We construct intra- and inter-modal hyperedges, which can connect an arbitrary number of nodes. In this fashion, DGLF enables the natural encoding of high-order relationships beyond pairwise formulation. **For the second issue**, we construct another vanilla graph to perform high-frequency enhanced propagation, by adopting a set of frequency filters (Dong et al., 2021; Bo et al., 2021), which distill different frequency constituents from node features. By adaptively integrating high-frequency enhanced messages, DGLF effectively captures sarcastic inconsistencies in local neighborhoods, which is vital for MSD. **Moreover**, we introduce multi-modal information bottleneck (Wu et al., 2023; Zhu et al., 2024a) to effectively fuse the learned graph features, which narrows down the solution space, driving the model's gaze toward shared modality information.

Overall, our contributions are three-fold: (1) We propose DGLF, a novel dual graph-based learning framework for MSD. To our best knowledge, we are the first to introduce hypergraph into MSD. (2) We construct a hypergraph and a vanilla graph to perform high-order aware and high-frequency enhanced propagation, respectively. Besides, we introduce multi-modal information bottleneck to effectively fuse the learned graph features. (3) Extensive experiments show that our model achieves new state-of-the-art results, further analyses confirm the effect of each component of our model.

2 Methodology

Task Definition. Given a sample $\mathbf{S}_i$ from the training set, the objective of MSD is to determine whether the sample implies any sarcasm by learning a model $f(\cdot)$ using the text $\mathbf{T}_i$ and corresponding image $\mathbf{V}_i$ of $\mathbf{S}_i$. This conventional training procedure is represented as $\hat{y}_i = f(\mathbf{T}_i, \mathbf{V}_i \,|\, \Theta) \in \{0, 1\}$, where $\hat{y}_i = 1$ indicates the sample is predicted to be sarcastic and vice versa; Θ represents the learnable model parameters. For simplicity, we temporally omit the superscript i that indexes the training samples in the following.

Next, before diving into the details of the proposed DGLF's architecture, we first introduce the feature encoding (§2.1) of modalities and construction of the dual graphs (§2.2) for the MSD task.

2.1 Feature Encoding

For a fair comparison with previous works, given a textual sentence $\mathbf{T} = \{t_1, t_2, \ldots, t_N\}$ consist-

ing of N words, we adopt the pre-trained BERT model (Devlin et al., 2019), to map each word t_* into d-dimensional embedding,[1] denoted as $\mathbf{H}^t \in \mathbb{R}^{N\times d}$. For a given image $\mathbf{V} \in \mathbb{R}^{L_h\times L_w}$, following Liang et al. (2021); Liu et al. (2022), we first resize it to 224×224 pixels, *i.e.*, $L = L_h = L_w = 224$. Then the image is divided into $M = p \times p$ patches[2], *w.r.t.* $\mathbf{V} \in \mathbb{R}^{M\times(L/p\times L/p)}$. Subsequently, we feed the sequence of M image patches into a Vision Transformer (ViT) (Dosovitskiy et al., 2021) with an MLP layer to acquire the visual representation $\mathbf{H}^v \in \mathbb{R}^{M\times d}$.

2.2 Graph Construction

2.2.1 Hypergraph Construction

After we obtain textual and visual representations $\mathbf{H}_t$ and $\mathbf{H}_v$ through §2.1, we construct a hypergraph $\mathcal{G}$ with edge-dependent node weights (Feng et al., 2019; Chitra and Raphael, 2019) from the representations as shown in Figure 3(a, b).

Mathematically, denote a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \tau, \zeta)$ in which each node $v_i^\delta \in \mathcal{V}(\delta \in \{t, v\}, |\mathcal{V}| = N + M)$ corresponds to a textual token v_i^t or visual token v_i^v, where we initialize the node embeddings $\{v_i^t, v_i^v\}$ with the encoded representations $\{h_i^t, h_i^v\}$, respectively.

For every hyperedge $e \in \mathcal{E}(|\mathcal{E}| = 2+N+M)$, it encodes intra- or inter-modal dependencies. Specifically, each node $v_i^\delta(\delta \in \{t, v\})$, where i spans the range 1 to N_δ with N_δ denoting the total number of tokens in modality δ, is first connected to all other tokens in the same modality $\{v_j^\delta | j \neq i, 1 \leq j \leq N_\delta\}$ through a single intra-modal hyperedge. Here we obtain 2 intra-modal hyperedges. Furthermore, each node v_i^δ is connected to all tokens in the opposite modality $\{v_k^{\bar{\delta}} | 1 \leq k \leq N_{\bar{\delta}}\}$, with $\bar{\delta}$ indicating the modality opposite to δ via an inter-modal hyperedge, where $N_{\bar{\delta}}$ represents the total number of tokens in modality $\bar{\delta}$. Here we obtain N and M inter-modal hyperedges from N textual tokens and visual tokens, respectively. The sum of $N + M$ inter-modal hyperedges and 2 intra-modal hyperedges results in $N + M + 2$ hyperedges.

Unlike prior methods (Liang et al., 2021, 2022; Liu et al., 2022) that resort to adjustments of edge weighting strategies through complex relationship learning or similarity metrics, our DGLF embraces simplicity by adopting randomly initialized weight values. Concretely, we introduce two distinct categories of weights in $\mathcal{G}$: (1) an edge weight $\tau(e)$ assigned to each hyperedge e, and (2) a node weight $\zeta_e(v)$ for every node v upon which hyperedge e is incident, denoted as edge-dependent node weight (Chitra and Raphael, 2019). $\zeta_e(v)$ quantifies the significance of node v within hyperedge e, thereby reinforcing fine-grained intra- and inter-modal relationships. Denote $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$ represent the incidence matrix, in which a nonzero entry $\mathbf{A}_{ve} = 1$ indicates that the hyperedge is incident with the node v; otherwise $\mathbf{A}_{ve} = 0$. Formally, edge-dependent node weights can be represented by a weighted incidence matrix $\hat{\mathbf{A}} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$:

$$\hat{\mathbf{A}} = \begin{cases} \zeta_e(v), & \text{if } e \text{ is incident with node } v, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

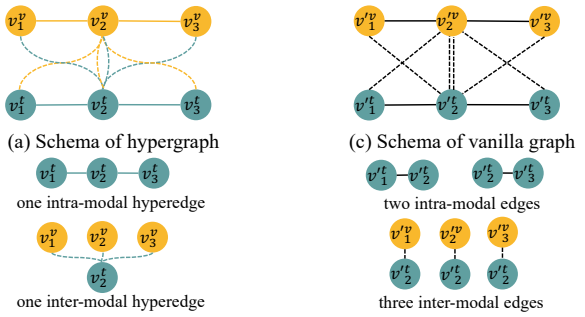

Figure 3: The illustration of hypergraph (a, b) and vanilla graph (c, d), *w.l.o.g*, only the edges directed into v_2^v, v_2^t, $v_2'^v$ and $v_2'^t$ are shown. As demonstrated, the hypergraph is adept at modeling high-order relationships, where four vertices in (a) (*e.g.*, v_1^v, v_2^v, v_3^v and v_2^t) are connected by a single hyperedge sharing the same color green; and in (c), the edge construction adheres to the pairwise relationship assumption.

2.2.2 Vanilla Graph Construction

As previously discussed, high-frequency information that reflects emotional discrepancies may be more pivotal for MSD (Bo et al., 2021; Wu et al., 2019), and combining the power of messages with different frequencies is worth exploring. This insight compels us to introduce a high-frequency enhanced propagation aimed at distilling varying frequency importance. To this end, we further construct a vanilla graph $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}'\}$ from the multi-modal encoded representations as shown in Figure 3(c, d), in parallel with the hypergraph. Mathematically, denote a vanilla graph $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}'\}$ whose nodes $\mathcal{V}'$ are identical to the ones in $\mathcal{G}$, denoted with $\{v_i'^t, v_j'^v\}$. The node embeddings are similarly initialized with the encoded

[1] Following Liang et al. (2021, 2022), we adopt the first sub-token's representation as the whole word representation.

[2] In line with previous works, p is set to 7.

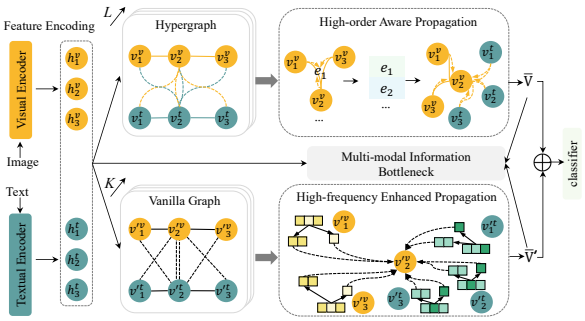

Figure 4: The architecture of DGLF, where dotted directed arrows in both propagation modules delineate the direction of message passing. Note that in High-frequency Enhanced Propagation component, ■ and ■ represent visual and textual high-frequency messages, while ■ and ■ represent visual and textual low-frequency messages. (*Zoom-in for better view*)

representations $\{h_i^t, h_j^v\}$ as well. The normalized graph Laplacian matrix can be represented as $\mathbf{L} = \mathbf{I} - \mathbf{D}_{\mathcal{G}'}^{-1/2}\mathbf{A}'\mathbf{D}_{\mathcal{G}'}^{-1/2}$, where $\mathbf{A}' \in \mathbb{R}^{|\mathbf{V}'|\times|\mathbf{V}'|}$ denotes the adjacency matrix, $\mathbf{D}_{\mathcal{G}'}$ denotes a diagonal degree matrix and $\mathbf{I}$ denotes an identity matrix.

2.3 Model Architecture

This section introduces the details of our proposed DGLF, whose architecture is shown in Figure 4.

2.3.1 High-order Aware Propagation

The main objective of constructing the hypergraph is to explore the subtle high-order sarcastic information within and across modalities. Concretely, we first conduct node convolution by aggregating node features to update hyperedge embeddings, and then conduct hyperedge convolution to spread the hyperedge messages to the nodes:

$$\mathbf{V}^{(\ell+1)} = \sigma(\mathbf{D}_{\mathcal{G}}^{-1}\mathbf{A}\mathbf{W}_e\mathbf{B}^{-1}\hat{\mathbf{A}}^\top\mathbf{V}^{(\ell)}), \tag{2}$$

in which $\mathbf{V}^{(\ell)} = \{v_{i,(\ell)}^{\delta} | \delta \in \{t, v\}, i \in [1, N] \text{ when } \delta = t; i \in [1, M] \text{ when } \delta = v\}$, $\mathbf{V}^{(\ell)} \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the input at layer ℓ; σ is a non-linear activation function; $\mathbf{W}_e = \text{diag}(\tau(e_1), \ldots, \tau(e_{|\mathcal{E}|})$ denotes the hyperedge weight matrix; $\mathbf{D}_{\mathcal{G}} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{B} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ denote the node degree and hyperedge degree matrix. In this manner, the high-order inter- and intra-modal relationships are gradually refined. After L iterations, we get the outputs of the last iteration $v_{i,(L)}^{\delta}$ as the high-order aware representations.

2.3.2 High-frequency Enhanced Propagation

To propagate high-frequency enhanced information on the vanilla graph, we first design a low-pass filter $\mathcal{F}_l$ and a high-pass filter $\mathcal{F}_h$ to distil the messages from the node features:

$$\begin{aligned} \mathcal{F}_l &= \mathbf{I} + \mathbf{D}_{\mathcal{G}'}^{-1/2}\mathbf{A}'\mathbf{D}_{\mathcal{G}'}^{-1/2} = 2\mathbf{I} - \mathbf{L}, \\ \mathcal{F}_h &= \mathbf{I} - \mathbf{D}_{\mathcal{G}'}^{-1/2}\mathbf{A}'\mathbf{D}_{\mathcal{G}'}^{-1/2} = \mathbf{L}. \end{aligned} \tag{3}$$

Note that $\mathcal{F}_h$ is equivalent to the normalized graph Laplacian matrix, which is consistent with the theory that the Laplacian kernel can be employed to highlight high-frequency information (Jain and Farrokhnia, 1991).

Now, we employ these two filters to adaptively aggregate messages with varying frequencies. To be specific, we use a weighted sum to combine low-frequency and high-frequency messages:

$$\begin{aligned} \mathbf{V}'^{(k+1)} &= \mathbf{W}^l(\mathcal{F}_l \cdot \mathbf{V}'^{(k)}) + \mathbf{W}^h(\mathcal{F}_h \cdot \mathbf{V}'^{(k)}) \\ &= \mathbf{V}'^{(k)} + (\mathbf{W}^l - \mathbf{W}^h)\mathbf{D}_{\mathcal{G}'}^{-1/2}\mathbf{A}'\mathbf{D}_{\mathcal{G}'}^{-1/2}\mathbf{V}'^{(k)}, \end{aligned} \tag{4}$$

where $\mathbf{V}'^{(k)} = \{v_{i,(k)}'^{\delta} | i \in [1, N] \text{ when } \delta = t; i \in [1, M] \text{ when } \delta = v, \delta \in \{t, v\}\} \in \mathbb{R}^{|\mathcal{V}'|\times d}$ is the input at layer k; $\mathbf{W}^l, \mathbf{W}^h \in \mathbb{R}^{|\mathcal{V}'|\times|\mathcal{V}'|}$ denote the weight matrices for low- and high-frequency information. It can be further rewritten as:

$$v'_{i,(k)+1} = v'_{i,(k)} + \sum_{j\in\mathcal{N}_i} \frac{w_{ij}^l - w_{ij}^h}{\sqrt{|\mathcal{N}_j|}\sqrt{|\mathcal{N}_i|}} v'_{j,(k)}, \tag{5}$$

where $\mathcal{N}_i$ denotes the neighboring nodes of node i; w_{ij}^l and w_{ij}^h denote the weight contributions of node j's low-frequency and high-frequency messages to node i, respectively with $w_{ij}^l + w_{ij}^h = 1$. And they are calculated using a self-gating mechanism similar to Bo et al. (2021).

By stacking K layers, each node receives the high-frequency enhanced messages, which are ignored by previous works, and we utilize outputs of the final layer $v_{i,(K)}'^{\delta}$ as the high-frequency enhanced representations.

2.3.3 Information Bottleneck based Multi-modal Fusion

Ideally, concatenated representations should encapsulate information shared across modalities. Thus, we introduce Multi-modal Fusion Information Bottleneck (MFB) (Wu et al., 2023; Mai et al., 2022), which effectively constrains the solution space to focus more on the shared multi-modal information. Concretely, we first fuse different modalities to obtain concatenated high-order aware and high-frequency enhanced representations as follows:

$$\bar{\mathbf{V}} = \mathbf{V}_{(L)}^t \oplus \mathbf{V}_{(L)}^v, \quad \bar{\mathbf{V}}' = \mathbf{V}_{(K)}'^t \oplus \mathbf{V}_{(K)}'^v, \tag{6}$$

Denote $\mathbf{H} = \mathbf{H}_t \oplus \mathbf{H}_v$, the MFB for concatenated high-order aware and high-frequency enhanced representations are calculated as follows:

$$\begin{aligned}\min_{p(\bar{\mathbf{V}}|\mathbf{H})\in\Omega_1} \text{MFB}(\bar{\mathbf{V}};\mathbf{H}) &\triangleq -\sum_{\delta}\mathcal{I}(\mathbf{H}_\delta;\bar{\mathbf{V}}) + \beta\mathcal{I}(\mathbf{H};\bar{\mathbf{V}}),\\ \min_{p(\bar{\mathbf{V}}'|\mathbf{H})\in\Omega_2} \text{MFB}(\bar{\mathbf{V}}';\mathbf{H}) &\triangleq -\sum_{\delta}\mathcal{I}(\mathbf{H}_\delta;\bar{\mathbf{V}}') + \beta\mathcal{I}(\mathbf{H};\bar{\mathbf{V}}'),\end{aligned} \tag{7}$$

where $\delta \in \{t, v\}$; β serves as a trade-off hyper-parameter; Ω_1 and Ω_2 denote the search space of the conditional distribution of $\bar{\mathbf{V}}$ and $\bar{\mathbf{V}}'$ given the initial concatenated vertex feature $\mathbf{H}$, respectively. We then perform bound estimations to enable its computation and training via back propagation:

Proposition 1. *The upper and lower bounds of mutual information between two random variables* $\mathbf{x}$ *and* $\mathbf{y}$ *can be estimated as:*

$$\mathbb{E}\left[\log \frac{f(\mathbf{y}_+,\mathbf{x})}{\sum_{\mathbf{y}_i\in Y} f(\mathbf{y}_i,\mathbf{x})}\right] \le \mathcal{I}(\mathbf{x};\mathbf{y}) \le D_{KL}(p(\mathbf{x}|\mathbf{y})||q(\mathbf{x})), \tag{8}$$

where $\mathbf{x}$ *and* $\mathbf{y}_+$ *are positive pairs sampled from* $p(\mathbf{x}|\mathbf{y})$*,* $f(\cdot,\cdot)$ *represents a scoring function that measures the similarity between two embeddings, and* q *represents a prior distribution of* $\mathbf{x}$*.*

The form of mutual information's lower bound above is known as the InfoNCE loss (Oord et al., 2018). Thus, the MFB loss can be expressed as :

$$\begin{aligned}\mathcal{L}_{\text{MFB}} = &\sum_{\delta}\mathcal{L}_{\text{InfoNCE}}(\mathbf{H}_\delta,\bar{\mathbf{V}}) + \beta D_{\text{KL}}(p(\bar{\mathbf{V}}|\mathbf{H})||q(\bar{\mathbf{V}}))\\ &+\sum_{\delta}\mathcal{L}_{\text{InfoNCE}}(\mathbf{H}_\delta,\bar{\mathbf{V}}') + \beta D_{\text{KL}}(p(\bar{\mathbf{V}}'|\mathbf{H})||q(\bar{\mathbf{V}}')).\end{aligned} \tag{9}$$

2.4 Training Objective

After the above procedures, we follow Liang et al. (2021, 2022) to employ attention mechanism (Vaswani et al., 2017) based on $\mathbf{H}$ and $\mathbf{V}$ (*resp.* $\mathbf{V}'$) to obtain the high-order aware presentation $\mathbf{f}_1$ (*resp.* high-frequency enhanced presentation $\mathbf{f}_2$). Now, we concatenate $\mathbf{f}_1$ and $\mathbf{f}_2$ to obtain the final representation $\mathbf{f}$, which is then fed into a fully-connected layer with softmax normalization to capture a probability distribution $\hat{\mathbf{y}} \in \mathbb{R}^{d_p}$ of sarcasm detection space as follows:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_o\mathbf{f} + \mathbf{b}_o), \tag{10}$$

where d_p is the dimensionality of sarcasm labels; $\mathbf{W}_o$ and $\mathbf{b}_o \in \mathbb{R}^{d_p}$ are trainable parameters. For MSD task, we minimize the cross-entropy loss via the standard gradient descent algorithm:

$$\mathcal{L}_{\text{MSD}} = -\sum_{i=1}^{N}\sum_{j=1}^{d_p}\mathbf{y}_i^j \log \hat{\mathbf{y}}_i^j + \lambda||\Theta||^2, \tag{11}$$

where N denotes the training data size; $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ represent the ground-truth and predicted label distribution of instance i, respectively; Θ denotes all trainable parameters of our DGLF and λ denotes the coefficient of L_2-regularization.

The final training objective is the sum of $\mathcal{L}_{\text{MSD}}$ (Eqn. (11)) and $\mathcal{L}_{\text{MFB}}$ (Eqn. (9)).

3 Experiments

3.1 Main Results

Due to space limitation, we put experiment settings in Appendix A.1. The performance comparison of our model and baselines are shown in Table 1, from which we have the following observations:

(1) ***Our model achieves new state-of-the-art (SOTA) performance on all metrics and datasets.*** Specifically, on MMSD, DGLF overpasses HKE by 1.62% and 2.91% on Acc. and F1; on MMSD2.0, it overpasses Att-BERT by 1.49% and 1.56% on Acc. and F1 respectively. This is because our model captures complicated high-order relationships based on the hypergraph, and our designed high-frequency enhanced propagation based on the vanilla graph further improves the model's ability to detect inconsistencies in sarcasm.

(2) ***The improvements on MMSD dataset are much sharper.*** This can be attributed to the fact that MMSD2.0 eliminated obvious sarcastic cues in MMSD, which has led to an obvious performance decline in existing GNN-based methods that rely on edge weighting strategies with complex relationship learning or similarity metrics, sometimes even underperforming compared to Att-BERT. Thanks to the high-order and high-frequency information captured, our model achieves consistent improvements over all baselines on both datasets.

(3) ***Based on more advanced feature encoders (e.g., CLIP), our DGLF can still achieve significant improvements.*** We suspect the reason is that the advantages of our approach are orthogonal to the ability of feature encoders. Our method can teach CLIP to model high-order relationships and high-frequency messages on dual graphs, which can hardly be learned in the pre-training process.

3.2 Method Analysis

Effect of High-order Aware Propagation. One of the core contributions of our work is modeling the high-order relationships based on the hypergraph, while previous GNN-based works only adhere to the pairwise relationship formulation. To

Modality	Model	MMSD				MMSD2.0			
		Acc. (%)	P (%)	R (%)	F1 (%)	Acc. (%)	P (%)	R (%)	F1 (%)
Text-only	TextCNN (Kim, 2014)	80.03	74.29	76.39	75.32	71.61	64.62	75.22	69.52
	Bi-LSTM (Graves and Schmidhuber, 2005)	81.90	76.66	78.42	77.53	72.48	68.02	68.08	68.05
	SMSD (Xiong et al., 2019)	80.90	76.46	75.18	75.82	75.36	68.45	71.55	69.97
	BERT* (Devlin et al., 2019)	83.60	78.50	82.51	80.45	76.52	74.48	73.09	73.78
Image-only	ResNet (He et al., 2016)	64.76	54.41	70.80	61.53	65.50	61.17	54.39	57.58
	ViT* (Dosovitskiy et al., 2021)	68.51	57.19	70.83	63.28	71.80	64.96	75.15	69.68
Multi-modal	HFM (Cai et al., 2019)	83.44	76.57	84.15	80.18	70.57	64.84	69.05	66.88
	D&R Net (Xu et al., 2020)	84.02	77.97	83.42	80.60	–	–	–	–
	Att-BERT (Pan et al., 2020)	86.05	80.87	85.08	82.92	80.03	76.28	77.82	77.04
	GNN-based, BERT & ViT as Encoder								
	InCrossMGs (Liang et al., 2021)	86.10	81.38	84.36	82.84	–	–	–	–
	CMGCN (Liang et al., 2022)	86.54	–	–	82.73	79.83	75.82	78.01	76.90
	HKE* (Liu et al., 2022)	87.39	81.40	86.93	84.07	76.39	73.50	75.96	74.71
	DGLF (Ours)	**89.01**†	**84.96**†	**89.10**†	**86.98**†	**81.52**†	**77.98**†	**79.23**†	**78.60**†
	CLIP as Encoder								
	Multi-view CLIP* (Qin et al., 2023)	88.22	82.03	88.13	84.97	85.14	80.18	88.21	84.00
	$\text{DGLF}_{\text{CLIP}}$ (Ours)	**89.43**†	**85.81**†	**89.27**†	**87.51**†	**86.82**†	**81.90**†	**89.85**†	**85.69**†

Table 1: Results comparison. $*$ denotes our re-implementation using the official code. - denotes missing results from the published work. Since CLIP-based methods use different pre-trained feature encoders, we gray out them for a fair comparison. † denotes the significance tests of DGLF and $\text{DGLF}_{\text{CLIP}}$ over baselines at p-value < 0.05.

verify its effect, we design variant 1 as shown in Table 2. We can observe that Acc. drops by 1.95% on MMSD and 1.65% on MMSD2.0. Moreover, F1 drops more significantly: 2.16% on MMSD and 1.81% on MMSD2.0. This proves that modeling the high-order relationships on the hypergraph $\mathcal{G}$ can naturally encode higher arity relationships within and between visual and textual elements, thus significantly improving sarcasm detection.

In §2.2.1, we define two types of weights in hypergraph $\mathcal{G}$ to capture the high-order relationships at a fine-grained level. To study its effect, we conduct the ablation experiments by removing different sets of them. It can be seen that from variant 2 to 4 in Table 2 that removing either or both (*i.e.*, setting weight value $\tau(e)$ or/and $\zeta_e(v)$ as 1) leads to performance decreases in all metrics on both datasets. This indicates that the formulated weights in hypergraph $\mathcal{G}$ benefit the final performance.

Effect of High-frequency Enhanced Propagation. The aim of constructing another vanilla graph $\mathcal{G}'$ is to perform high-frequency enhanced propagation. To verify its effectiveness, we design variant 5 by performing predictions using the high-order aware propagation on the hypergraph only. From Table 2, we find that variant 5 obtains dramatic drops in all metrics on both datasets. The apparent performance gap verifies the high-frequency enhanced messages in MSD, which can capture the varying importance of sentiment discrepancy and commonality within local neighborhoods.

To further analyze the propagation of only high-frequency and low-frequency information in the vanilla graph $\mathcal{G}'$, we set $\mathbf{R}^l$ and $\mathbf{R}^h$ in Eqn. (4) to zero, respectively. From variant 6 and variant 7 in Table 2, we observe that erasing high-frequency information results in a more noticeable performance decline. This fully demonstrates the necessity of high-frequency information for the MSD task and intuitively supports our motivation.

Effect of Multi-modal Information Bottleneck. From Table 2 variant 8, we can find that removing $\mathcal{L}_{\text{MFB}}$ in final training loss leads to large performance decreases. Specifically, Acc. drops 1.28% and 0.56% on MMSD and MMSD2.0; F1 drops 1.37% and 0.67% on MMSD and MMSD2.0. This can verify the effectiveness of MFB, which constrains the solution space of the two learned graph representations to focus on cross-modal shared information, thus boosting performance.

Sensitivity of Hyper-parameter. To investigate the robustness of our proposed multi-modal information bottleneck, we conduct sensitivity analyses.

#	Variant	MMSD		MMSD2.0	
		Acc. (%)	F1 (%)	Acc. (%)	F1 (%)
-	DGLF	**89.01**	**86.98**	**81.52**	**78.60**
	Effect of High-order Aware Propagation				
1	w/o high-order aware propagation	87.06 (↓1.95)	84.82 (↓2.16)	79.87 (↓1.65)	76.79 (↓1.81)
2	w/o edge weight $\tau(e)$	87.25 (↓1.76)	85.17 (↓1.81)	80.81 (↓0.71)	77.80 (↓0.80)
3	w/o node weight $\zeta_e(v)$	87.81 (↓1.20)	85.69 (↓1.29)	80.98 (↓0.54)	77.97 (↓0.63)
4	w/o both weights	87.08 (↓1.93)	84.86 (↓2.12)	80.35 (↓1.17)	77.31 (↓1.29)
	Effect of High-frequency Enhanced Propagation				
5	w/o high-frequency enhanced propagation	86.73 (↓2.28)	84.17 (↓2.81)	79.51 (↓2.01)	76.28 (↓2.32)
6	w/o low-frequency messages included	88.07 (↓0.94)	85.86 (↓1.12)	80.67 (↓0.85)	77.59 (↓1.01)
7	w/o high-frequency messages included	87.84 (↓1.17)	85.60 (↓1.38)	80.46 (↓1.06)	77.37 (↓1.23)
	Effect of Multi-modal Information Bottleneck				
8	w/o multi-modal information bottleneck	87.73 (↓1.28)	85.61 (↓1.37)	80.96 (↓0.56)	77.93 (↓0.67)

Table 2: Results of ablation experiments.

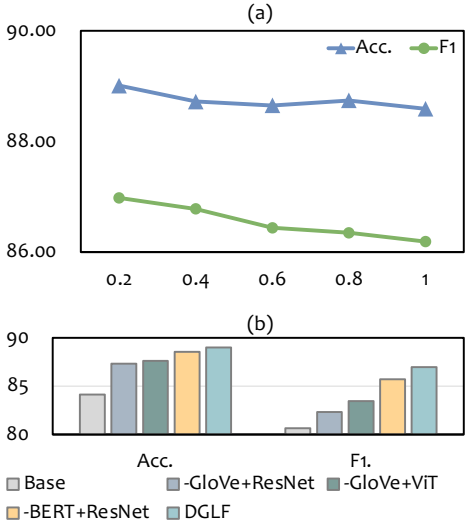

Figure 5: (a) Performance variations when altering the value of β in MFB loss (Eqn. (9)) on MMSD dataset. (b) Performance of using different pre-trained methods on MMSD. **Base** denotes without the proposed dual graphs, *i.e.*, only using BERT and ViT to conduct MSD.

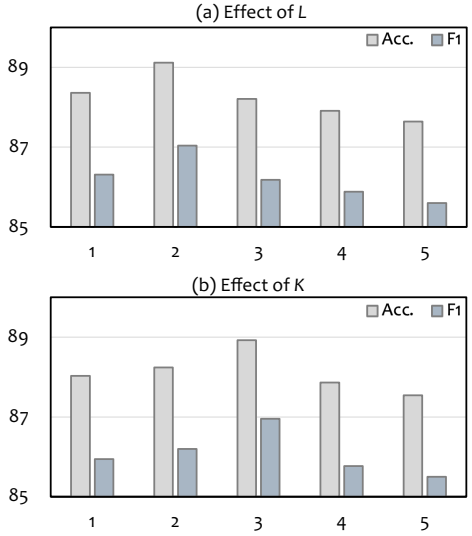

Figure 6: Results of the proposed DGLF at different graph layers (a) L and (b) K on MMSD dataset.

We varied the value of β from 0.2 to 1.0 as per Eqn. (9) and the results are displayed in Figure 5(a). As depicted, the performance remains relatively stable across different values of β, albeit with a slight decreasing trend as β increases. This suggests that our model is largely insensitive to β. However, as β increases and the constraint tightens, there is a gradual effect on the model performance.

Impact of Graph Layers. To investigate the impact of stacking different graph layers for hypergraph $\mathcal{G}$ and vanilla graph $\mathcal{G}'$, we conduct a grid search on the number of layers L and K, respectively. Specifically, we search them from 1 to 5 on the validation set and the results on MMSD dataset are shown in Figure 6. We observe that the effects of L and K are similar. At first, the results steadily improve as stacking more layers, and peak at $L = 2$ and $K = 3$ respectively. Further stacking more layers has little positive impact, as it may incorporate noise from neighborhoods. For MMSD2.0, we empirically found that the performance is not particularly sensitive to the number of layers, displaying no specific pattern.

Generalizability of Dual Graphs. To evaluate the generalizability of the proposed dual graphs

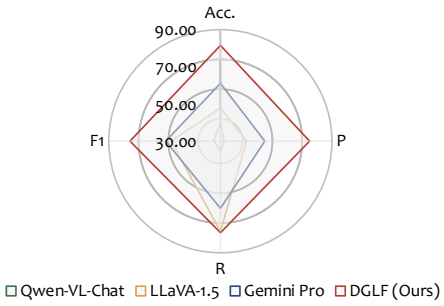

Figure 7: Comparison with large vision-language models (LVLMs) on MMSD2.0 dataset.

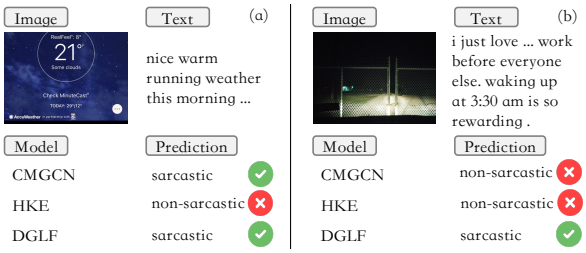

Figure 8: Case study.

with various pre-trained methods beyond CLIP, we conduct experiments using four variants that employ different textual and visual encoders. From results in Figure 5(b), we find that the dual graphs are compatible with various pre-trained models and perform consistently outperform the baseline model without graphs. This verifies the effectiveness and generalizability of the proposed DGLF. Further, it is evident that using more sophisticated pre-trained methods, such as ViT, BERT and CLIP, leads to superior performance.

Comparison with Large Vision-Language Models (LVLMs). To ascertain the competitive edge of our model, we conducted comparative analyses against prevalent LVLMs including Qwen-VL-Chat (Bai et al., 2023), LLaVA-1.5 (Liu et al., 2024), and Gemini Pro (Team et al., 2023) following Wang et al. (2024). The obvious performance gap between ours and LVLMs in Figure 7 underscores the persistent challenges that LVLMs encounter in MSD, despite their advanced zero-shot learning and chain-of-thought capabilities. It emphasizes the need for dedicated efforts in designing effective MSD frameworks. Combining the power of LVLMs (Chen et al., 2024) to detect sarcasm is an interesting direction in future work.

3.3 Case Study

We present two sarcastic cases in Figure 8 to quantitatively verify the effectiveness of our DGLF.

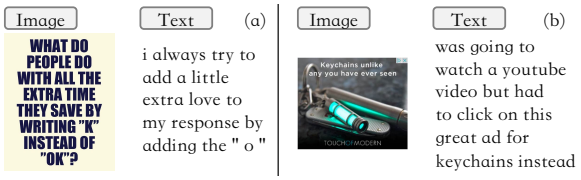

Figure 9: Examples of error prediction.

Model	Acc. (%)	P (%)	R (%)	F1 (%)
DGLF	81.52	77.98	79.23	78.60
+ OCR	81.64	77.10	81.06	79.03

Table 3: Results on incorporating Optical Character Recognition (OCR) information on MMSD2.0 dataset.

Case (a). The sarcasm arises from the contrast between the text, which suggests pleasant and warm conditions, and the image, which displays a much colder temperature typically not considered warm. Despite involving a more complex fusion architecture, HKE still employs a naive GNN propagation framework, transmitting low-frequency messages between nodes. In contrast, our model obtains high-order and high-frequency messages from both the hypergraph and the vanilla graph, and accurately captures the inter-modal inconsistencies, thereby correctly identifying it as sarcastic.

Case (b). The sarcasm likely stems from the exaggeration between the text and the image. The text describes waking up extremely early as "*rewarding*", which contrasts sharply with the image accompanying it, depicting the reality of driving early in the morning. Unfortunately, previous SOTA GNN-based methods still predict it as non-sarcastic. Thanks to the proposed dual graph framework, our model can propagate high-frequency messages that reflect discrepancies in the graph, thereby effectively capturing inter-modality inconsistencies and achieving precise prediction.

3.4 Error Analysis

We further conduct error analysis to understand DGLF's performance. We observe that the majority of errors occur in samples where images contain important textual information, such as the purely textual image in Figure 9 (a) and in Figure 9 (b) which includes both textual and visual expressions. Based on these observations, we conducted a preliminary experiment to leverage textual information within images by integrating OCR into DGLF for MSD. From the results in Table 3, we find a no-

table performance improvement. Thus it would be interesting to effectively leverage important textual information in images for future research.

4 Related Work

Multi-modal Sarcasm Detection. With the rapid popularization of social media platforms, multi-modal sarcasm detection (MSD) has garnered increasing research attention in recent years (Zhu et al., 2024c; Wang et al., 2024). Some early works (Xu et al., 2020; Pan et al., 2020; Xin et al., 2024) focused more on contextual dependencies and utilized feature concatenation for multi-modal modeling. More recently, researchers formulated the MSD task upon GNNs, which have shown promising results. Therein, InCrossMGs (Liang et al., 2021) proposed in-modal and cross-modal graphs to determine the sentiment inconsistencies. Based on this, CMGCN (Liang et al., 2022) explored a cross-modal graph to model the contradictory sentiments between key textual and visual information. HKE (Liu et al., 2022) further introduced a GNN-based hierarchical framework by exploring both the atomic-level congruity and the composition-level congruity.

Nevertheless, these GNN-based MSD models still deliver insufficient high-order relationships and high-frequency messages, as we discussed.

Graph Neural Networks. GNNs can explicitly model data relationships, which have been widely employed in various applications such as sentiment analysis and argument pair extraction (Li et al., 2021; Chen et al., 2023; Sun et al., 2023; Zhu et al., 2024b). GNNs have also inspired MSD researchers and offer a new solution for the MSD task, from unimodal setting (Lou et al., 2021) to multi-modal scenario (Qin et al., 2023).

However, previous works fail to address the general limits of GNNs, which motivates our work.

Multi-modal Information Bottleneck. The InfoMax principle proposed by Linsker (1988) seeks to maximize the mutual information between feature and model output. Along this way, Han et al. (2021) built up a hierarchical mutual information maximization guided model for multi-modal sentiment analysis. Wu et al. (2023) focused on video-based sentiment analysis and used contrastive learning to achieve mutual information maximization.

In this work, we utilize the lower bound of multi-modal information bottleneck (Mai et al., 2022) to constrain the learned graph features and initial modality features, driving the model's gaze toward shared modality information.

5 Conclusion

In this paper, we propose a new GNN-based framework for MSD. We construct a hypergraph and a vanilla graph to perform high-order aware propagation and high-frequency enhanced propagation, respectively. Based on this, we introduce multi-modal information bottleneck to effectively fuse the two learned graph representations. Extensive experiments and analyses on two MSD benchmarks show the superiority of our proposed framework.

Limitations

Our DGLF has the following limitations: (1) The proposed dual graph approach (combining hypergraph and vanilla graph) might lead to increased computational complexity. (2) The effectiveness of the DGLF might be contingent upon the quality of the underlying pre-trained models (BERT, ViT), and can benefit from more advanced feature encoders, which is not the focus of this work. (3) DGLF lacks validation on more diverse multi-modal MSD datasets that include additional modalities such as audio and video, as well as testing across various other tasks, which may limit its broader generalizability and effectiveness.

Acknowledgments

This work was supported by the Major Program of the National Natural Science Foundation of China (No. 61991410).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Song Bai, Feihu Zhang, and Philip HS Torr. 2021. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637.

Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. 2021. Beyond low-frequency information in graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3950–3957.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical

fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515. Association for Computational Linguistics.

Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10761–10770.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. 2024. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*.

Uthsav Chitra and Benjamin Raphael. 2019. Random walks on hypergraphs with edge-dependent vertex weights. In *International conference on machine learning*, pages 1172–1181. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yushun Dong, Kaize Ding, Brian Jalaian, Shuiwang Ji, and Jundong Li. 2021. Adagnn: Graph neural networks with adaptive frequency response filter. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 392–401.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3558–3565.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Anil K Jain and Farshid Farrokhnia. 1991. Unsupervised texture segmentation using gabor filters. *Pattern recognition*, 24(12):1167–1186.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4707–4715. ACM.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1767–1777. Association for Computational Linguistics.

Ralph Linsker. 1988. Self-organization in a perceptual network. *Computer*, 21(3):105–117.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006. Association for Computational Linguistics.

Chenwei Lou, Bin Liang, Lin Gui, Yulan He, Yixue Dang, and Ruifeng Xu. 2021. Affective dependency graph for sarcasm detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1844–1849.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1383–1392. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1601–1612.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2.0: towards a reliable multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10834–10845. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Xiangguo Sun, Hongzhi Yin, Bo Liu, Hongxu Chen, Jiuxin Cao, Yingxia Shao, and Nguyen Quoc Viet Hung. 2021. Heterogeneous hypergraph embedding for graph classification. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 725–733.

Yang Sun, Bin Liang, Jianzhu Bao, Yice Zhang, Geng Tu, Min Yang, and Ruifeng Xu. 2023. Probing graph decomposition for argument pair extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13075–13088, Toronto, Canada. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Peng Wang, Yongheng Zhang, Hao Fei, Qiguang Chen, Yukai Wang, Jiasheng Si, Wenpeng Lu, Min Li, and Libo Qin. 2024. S3 agent: Unlocking the power of vllm for zero-shot multi-modal sarcasm detection. *ACM Trans. Multimedia Comput. Commun. Appl.* Just Accepted.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR.

Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. 2023. Denoising bottleneck with mutual information maximization for video multimodal fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2231–2243, Toronto, Canada. Association for Computational Linguistics.

Yifei Xin, Xuxin Cheng, Zhihong Zhu, Xusheng Yang, and Yuexian Zou. 2024. Diffatr: Diffusion-based generative modeling for audio-text retrieval. In *Interspeech 2024*, pages 1670–1674.

Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The World Wide Web Conference*, WWW '19, page 2115–2124, New York, NY, USA. Association for Computing Machinery.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 3777–3786, Online. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers*, pages 2449–2460.

Zhihong Zhu, Xuxin Cheng, Zhaorun Chen, Yuyan Chen, Yunyan Zhang, Xian Wu, Yefeng Zheng, and Bowen Xing. 2024a. Inmu-net: Advancing multi-modal intent detection via information bottleneck and multi-sensory processing. In *ACM Multimedia 2024*.

Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024b. Dance with labels: Dual-heterogeneous label graph interaction for multi-intent spoken language understanding. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1022–1031, New York, NY, USA. Association for Computing Machinery.

Zhihong Zhu, Xianwei Zhuang, Yunyan Zhang, Derong Xu, Guimin Hu, Xian Wu, and Yefeng Zheng. 2024c. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6687–6695. International Joint Conferences on Artificial Intelligence Organization. Main Track.

A Appendix

A.1 Experimental Settings

Datasets. We conduct experiments on two widely-used benchmark datasets: MMSD (Cai et al., 2019) and MMSD2.0 (Qin et al., 2023). Specifically, MMSD is derived from English Twitter. Thereinto, tweets with some special hashtags (*e.g.*, sarcasm) are positive examples and those without such hashtags are negative examples. MMSD2.0 is updated from MMSD, which is the most advanced benchmark in MSD. The statistics of these two datasets are shown in Table 4.

Evaluation Metrics. Following previous works (Liu et al., 2022; Qin et al., 2023), we adopt accuracy (Acc.), precision (P), recall (R), and F1 score (F1) to evaluate the model performance.

Implementation Details. For a fair comparison, we follow Liang et al. (2021, 2022); Liu et al. (2022) to utilize the pre-trained uncased BERT-base model (Devlin et al., 2019) to embed each word as a 768-dimensional embedding and employ the pre-trained ViT (Dosovitskiy et al., 2021) to embed each visual patch as a 768-dimensional embedding, *i.e.*, $d = 768$. Adam (Kingma and Ba, 2014) is utilized as the optimizer with a learning rate of $2e$-5, and the mini-batch size is 16. The coefficient λ is set to $1e$-5. We test L and K in the range from 1 to 5 on the validation set and choose the best-performing one to the test set, respectively. The hyper-parameter β is set as 0.2. Paired t-test is performed to test the significance of performance improvement with a default significance level of 0.05. All experiments are conducted on one NVIDIA GeForce RTX 3090. The results reported in all experiments are averages of 5 runs with different random seeds to ensure the final reported results are statistically stable.

MMSD/MMSD2.0	Train	Validation	Test
Sentences	19,816/19,816	2,410/2,410	2,409/2,409
Positive	8,642/9,572	959/1,042	959/1,037
Negative	11,174/10,240	1,451/1,368	1,450/1,372

Table 4: Statistics of two experimental datasets.

A.2 Model Zoo

We compared our proposed DGLF with a series of strong baselines, which can be broadly classified into three main categories:

(1) Text-only methods. These methods purely rely on textual information for sarcasm detection, including **TextCNN** (Kim, 2014), a deep learning model based on CNN; **Bi-LSTM** (Graves and Schmidhuber, 2005), a bidirectional LSTM network for text classification; **SMSD** (Xiong et al., 2019) explored a self-matching network to capture textual incongruity information; and **BERT** (Devlin et al., 2019), the vanilla pre-trained uncased BERT-base taking '`[CLS] text [SEP]`' as input. **(2) Image-only methods**. The sarcasm detection in these methods relies solely on image input, including **ResNet** (He et al., 2016) which trains a sarcasm classifier; and **ViT** (Dosovitskiy et al., 2021), which utilizes the '`[class]`' token representations to detect the sarcasm. **(3) Multi-modal methods**. These methods utilize both visual and textual information for sarcasm detection, including **HFM** (Cai et al., 2019) introduced a hierarchical fusion model for MSD; **D&R Net** (Xu et al., 2020) proposed a decomposition and relation network modeling both cross-modality contrast and semantic association; **Att-BERT** (Pan et al., 2020) explored inter-modality attention and co-attention

to model the incongruity of multimodal information; **InCrossMGs** (Liang et al., 2021) leveraged the sarcasm relations from both intra- and inter-modality perspectives using local multi-modal features; **CMGCN** (Liang et al., 2022) explored the sarcastic relations across objects of the image and tokens of the text; **HKE** (Liu et al., 2022) utilized both the atomic-level congruity based on cross attention and the composition-level congruity based on GNNs; and **Multi-view CLIP** (Qin et al., 2023) employed the pre-trained CLIP (Radford et al., 2021) model to detect different sarcastic cues captured from multiple perspectives.

Further, to investigate the effectiveness of our DGLF when used with different pre-trained models, we also set the following variants: **-GloVe+ViT**: We replace pre-trained BERT in our proposed framework with GloVe (Pennington et al., 2014) to initialize each word into a 300-dimensional embedding and utilize ViT for learning image-modality representations. **-BERT+ResNet**: Following Pan et al. (2020), we replace the ViT in our framework with ResNet-152 (He et al., 2016) to embed each image patch as a 2048-dimensional vector. **-GloVe+ResNet**: We use GloVe to acquire word embeddings and employ ResNet for learning image-modality representations.

A.3 Proof of Proposition 1

Proof. The proof of mutual information's lower bound estimation can be found in the appendix of Oord et al. (2018). Here we present the proof for the upper bound estimation. We know that the KL divergence is always greater than zero, and therefore we have:

$$D_{KL}(p(\mathbf{x})||q(\mathbf{x})) = \mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{x})] - \mathbb{E}_{q(\mathbf{x})}[\log q(\mathbf{x})] \geq 0. \tag{12}$$

By following the definition of mutual information, we get:

$$\begin{aligned} \mathcal{I}(\mathbf{x};\mathbf{y}) &= \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[\log \frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})}] \\ &\approx \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[\log \frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})}] \\ &\leq \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[\log \frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x}|\mathbf{y})}] \\ &= D_{KL}(p(\mathbf{x}|\mathbf{y})||q(\mathbf{x})). \end{aligned} \tag{13}$$

Thus, we conclude:

$$\mathcal{I}(\mathbf{x};\mathbf{y}) \leq D_{KL}(p(\mathbf{x}|\mathbf{y})||q(\mathbf{x})). \tag{14}$$

□