

# Cross-Modal Causal Scheduling for Enhancing Target-Oriented Multi-Modal Sentiment Classification

Pengyu Zhao<sup>1\*</sup>, Chaoyang Li<sup>1,2\*</sup>, Lingzhi Wang<sup>1</sup>, and Qing Liao<sup>1,2</sup> (✉)

<sup>1</sup> Harbin Institute of Technology, Shenzhen, China

<sup>2</sup> Peng Cheng Laboratory, Shenzhen, China

{23S151160, 22b951022}@stu.hit.edu.cn, {wanglingzhi, liaoqing}@hit.edu.cn

**Abstract.** Target-oriented multi-modal sentiment classification (TMSC) aims to identify sentiment polarity towards specific targets by considering multiple modalities, e.g., text and images. However, current methods often ignore spurious correlations within the data, which can cause models to learn irrelevant features that misrepresent the sentiment of targets. To address this issue, we propose a novel *Cross-Modal Causal Scheduling* framework (CMCS) that prioritizes learning multi-modal features with fewer spurious correlations. Specifically, we first design a *Multi-modal Feature Selection* model (MFS) that utilizes causal intervention to select relevant features. Second, we construct a *Causal cross-Modal Scheduler* (CMS) to assess the causal effects of selected features, which further optimize the multi-modal learning process based on these effects. Finally, we formulate the CMS and the multi-modal learning process as a bi-level optimization problem. In the lower optimization, the MFS is updated with the scheduled gradient, while in the upper optimization, the CMS is updated with the implicit gradient. Extensive experiments demonstrate that our method outperforms existing baseline methods on TMSC and can effectively schedule the learning process of multi-modal features based on causal effects.

**Keywords:** Multi-modal Sentiment Analysis · Target-oriented multi-modal sentiment classification · Causal Inference .

## 1 Introduction

Target-oriented multi-modal sentiment classification (TMSC) [27, 19, 47] is a challenging fine-grained sentiment analysis task that determines the sentiment polarity of opinion targets by considering various modalities, such as text and images. Taking Fig. 1 (a) as an example, in the sentence “*Pat celebrating her 90th birthday with Emily Roux in Chez Roux at the Newmarket Guineas Festival*”, three distinct targets can be identified: “*Pat*”, “*Emily Roux*”, and “*Newmarket Guineas Festival*”. The corresponding sentiment polarities for these targets are “*positive*”, “*positive*”, and “*neutral*”, respectively. TMSC has gained notable attention in multi-modal sentiment analysis due to the challenges of simultaneously handling different modalities [24, 11, 53]. Most existing works on TMSC primarily focus on effectively fusing multi-modal information,

---

\* The first two authors contributed equally to this work.

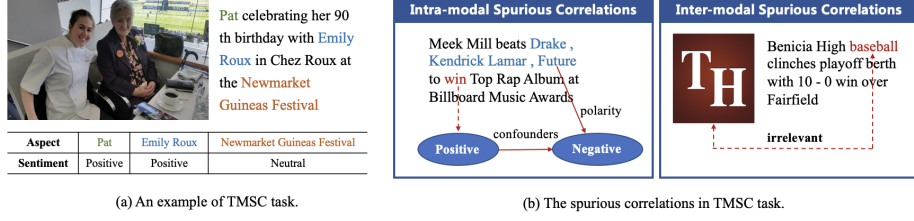


Fig. 1: The example and spurious correlations in the TMSC task. In (b), the left side depicts intra-modal spurious correlations, where the sentiment of the blue target may be influenced by the word “win”. The right side shows inter-modal spurious correlations, where the irrelevant visual features may interfere with the sentiment of the red target.

including methods like feature concatenation [25, 41], cross-modal alignment of image regions with text sequences [47, 11, 55], and using Energy-Based Models [30, 15] to enhance TMSC performance.

Despite progress in the TMSC field, most existing works overlook the spurious correlations within the multi-modal data, easily learning features irrelevant to the sentiment polarities of targets and degrading TMSC performance [37]. On the one hand, for a text sentence, a correlation bias (i.e., intra-modal spurious correlation) often exists between targets and co-occurring contextual words, which may lead the model to focus on words irrelevant to the sentiment of the targets, harming sentiment classification performance [54]. As shown in Fig. 1 (b), the word “win” is typically associated with positive sentiment, which may interfere with the negative sentiment polarity of targets (i.e., blue-marked words). On the other hand, given text-image pairs, images often contain information irrelevant to the text, leading models to erroneously associate irrelevant visual features with sentiment labels during training (i.e., inter-modal spurious correlations) [52]. As shown in Fig. 1 (b), the text “Benicia High baseball clinches a playoff berth with 10-0 win over Fairfield” lacks meaningful correlation with its corresponding image, which can cause the model to learn irrelevant visual features mistakenly.

To identify the causes of spurious correlations and suggest solutions, we construct a Structural Causal Model (SCM) (Fig. 2), where  $T$ ,  $I$ ,  $C$  and  $Y$  represent the text sentence, image, confounding factors and the sentiment predictions of targets, respectively. Here,  $T \rightarrow Y$  and  $I \rightarrow Y$  denote the desired causal effect, enabling the model to predict label  $Y$  directly from image  $T$  and  $I$ . However, not all features in  $T$  and  $I$  are relevant to  $Y$ . The confounding factors  $C$ , stemming from data bias, can interfere between  $T$  and  $Y$ , as well as between  $I$  and  $Y$ , creating spurious correlations with irrelevant features [13]. To

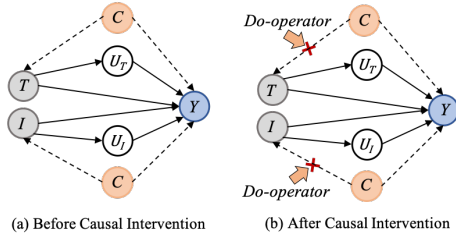


Fig. 2: The SCMs of TMSC.  $T$ ,  $I$ ,  $C$ ,  $U_T$ ,  $U_I$ ,  $Y$  denote text, image, confounders, textual/visual features, and predictions. Solid and dotted arrows indicate causal relationships and spurious correlations.  $Do$  – operator is an intervention operation.

address this, we aim to mitigate spurious correlations via causal intervention (i.e., *Do* – operator), extracting relevant textual features  $U_T$  and relevant visual features  $U_I$  to learn beneficial causal effects (i.e.,  $T \rightarrow U_T \rightarrow Y$  and  $I \rightarrow U_I \rightarrow Y$ ). Notely, different modal features can introduce varying degrees of spurious interference on sentiment labels, with samples exhibiting strong spurious correlations presenting greater challenges for model learning and hindering the convergence of multi-modal feature learning. This motivates us to explore a method that adaptively optimizes the learning process of multi-modal features based on causal effects, ultimately enhancing TMSC performance.

To achieve the above goal, we propose a novel *Cross-Modal Causal Scheduling* framework (CMCS), which prioritizes the learning of multi-modal features based on their susceptibility to spurious correlations. Our framework consists of three main components. First, we introduce the *Multi-modal Feature Selection* model (MFS), which leverages causal interventions to identify key features across modalities. Second, we design the *Causal cross-Modal Scheduler* (CMS), which employs counterfactual reasoning to assess the causal effects of selected features and schedule them for learning accordingly. Lastly, we implement a bi-level optimization strategy that determines the optimal cross-modal scheduling. The lower-level optimization updates the MFS using the scheduled gradient, while the upper-level optimization updates the CMS by implicit gradient updates. The key contributions are summarized as follows:

- We propose a novel *Cross-Modal Causal Scheduling* framework (CMCS) that optimizes multi-modal feature learning from a causal perspective, enhancing the performance of TMSC.
- We design a multi-modal feature selection model, which selects relevant multi-modal features with targets by simple yet effective causal intervention, alleviating spurious correlations.
- We construct a causal cross-modal scheduler that manages multi-modal learning processes by assessing the causal effects of features, incorporating a bi-level optimization to determine the optimal scheduling adaptively.
- Extensive experiments on two benchmark datasets show the superiority of our proposed framework over several baselines in terms of TMSC.

## 2 Related Work

### 2.1 Target-oriented Sentiment Classification

Sentiment analysis, also known as opinion mining, is a key research area in natural language processing and data mining [20, 45]. It focuses on systematically identifying affective states in textual data, enabling computational evaluation of emotions, opinions, and attitudes in written or spoken language [9].

Over the past decade, Target-Oriented Sentiment Classification (TSC), also known as Aspect-Based Sentiment Analysis (ABSA) [21], has become a key subfield of sentiment analysis, primarily focusing on identifying the sentiment polarity of target words

in textual data [4, 26]. Early studies in target-oriented sentiment classification primarily focused on modeling the structural relationships between target words and their contextual environments. For instance, introducing multi-grained attention architectures [6] to explicitly capture fine-grained linguistic dependencies between target words and their surrounding context. Others leverage graph convolutional networks, utilizing syntactic dependency trees and semantic role labeling frameworks to represent the bidirectional interdependencies between target words and their contextual descriptors [17, 16, 28].

More recently, causal inference has been incorporated into target-oriented sentiment classification to address issues related to data biases and spurious correlations between target words and their context. For example, some studies have introduced Structural Causal Models (SCMs) to disentangle confounding factors [54], while others have employed prompt-enhanced Large Language Models (LLMs) to generate counterfactually augmented training samples, effectively mitigating dataset biases [38].

## 2.2 Target-oriented Multi-modal Sentiment Classification

In recent years, with advancements in multimedia technology, social media data has exhibited a multi-modal trend, leading to widespread research interest in multi-modal sentiment analysis [51, 1, 11]. This has led to growing interest in target-oriented multi-modal sentiment classification (TMSC), which seeks to utilize both visual and textual content for more accurate sentiment predictions.

Most works in TMSC focus on fusing multi-modal information to improve sentiment analysis accuracy. For example, [2] bridges the gap between text and images using image attributes, while ESAFN [48] applies LSTM for entity-level sentiment analysis. [49] leverages BERT [5] for aspect-sensitive representations, offering deeper insights into sentiment in specific contexts. Additionally, [12] translates images into text to improve sentiment prediction, and [39] introduces multi-modal retrieval to refine text-image integration. [44] translates facial expressions into emotional semantics, connecting visual cues to emotional understanding. [18] advances vision-language pre-training for richer modality representations. [55] introduces an aspect-aware attention module that enhances the model’s ability to focus on relevant features tied to specific sentiment aspects. [50] captures image-target relationships. [30] integrates Energy-Based Models [15] into TMSC, refining the fusion process and boosting sentiment classification by modeling energy-based relationships between modalities.

**Differences.** Existing works primarily focus on fusion and alignment across modalities, often overlooking spurious correlations within multi-modal data. In contrast, we propose a novel cross-modal causal scheduling framework that extracts multi-modal features and assigns weights based on causal effects, reducing spurious correlations.

## 2.3 Causal Inference

Causal inference has gained attention for improving predictions by focusing on causal relationships rather than mere statistical correlations. For instance, structural causal models and do-calculus formalize causal relationships and guide interventions [29]. Neural networks can estimate causal effects between input variables and output targets [34, 10]. [8] mitigate the spurious association problem in the sarcasm detection

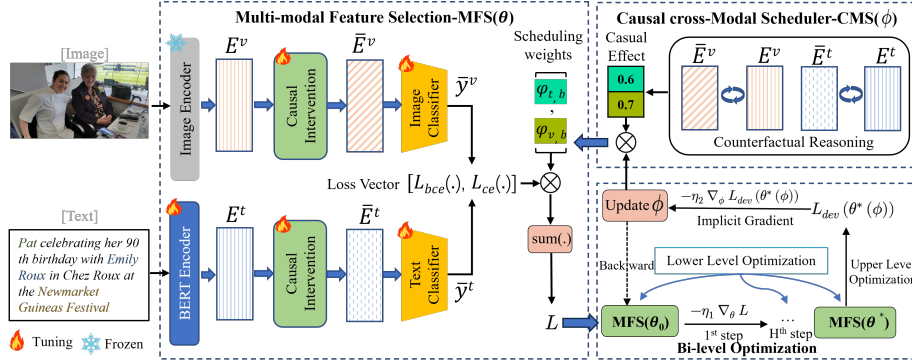


Fig. 3: An overview of CMCS training. First, the MFS selects relevant features with targets via causal interventions. These selected features are further input into classifiers for sentiment prediction. Second, the CMS assesses the causal effects of features using counterfactual reasoning, dynamically assigning learning weights based on causal effects. Last, the bi-level optimization is to solve optimal scheduling weights for enhanced multi-modal feature learning.

task through causal interventions. [46] alleviate the problem of spurious correlations of visual modalities through causal interventions. Although few works apply causal inference to sentiment analysis, such as [54] for text-aspect-based sentiment and [36, 35, 43] for traditional multi-modal sentiment, research on causal inference in TMSC is limited. Our method not only measures the causal effects of different modal features but also adaptively schedules them based on these effects, a consideration missing in existing approaches.

### 3 Methodology

#### 3.1 Overview.

**Task Definition.** Given a set of multi-modal samples  $\mathcal{X}$ , each sample  $X \in \mathcal{X}$  contains a text sentence  $X_t$  with  $n$  words and its corresponding image  $X_v$ , as well as the  $l$  opinion targets  $\mathcal{T} = (T_1, T_2, \dots, T_l)$  referring to a span in the sentence  $X_t$ . TMSC aims to predict the sentiment label  $y$  of each opinion target mentioned in the text-image pair  $X = (X_t, X_v)$ , where  $y$  can be either positive, negative, or neutral.

**Our framework.** Fig. 3 illustrates our proposed CMCS training framework. Specifically, to reduce the interference from spurious features, we first introduce the *Multi-modal Feature Selection* model (MFS), employing causal intervention to select relevant multi-modal features with targets. These selected features are then fed into the classifiers to generate predictions. Considering that different modal features exhibit varying levels of spurious interference with sentiment labels, we also develop a *Causal cross-Modal Scheduler* (CMS) that uses counterfactual reasoning to assess the causal effects of selected features and assigns learning weights accordingly. Finally, bi-level optimization is implemented to adaptively find the optimal weights for scheduling multi-modal learning.

### 3.2 Multi-modal Feature Selection

Spurious correlations may mislead the model into learning irrelevant textual and visual features to sentiment labels, further degrading TMSC effectiveness. To address these issues, we propose the *Multi-modal Feature Selection* model (MFS) that utilizes simple yet effective causal intervention to identify relevant features with targets' labels.

**Textual Feature Selection.** Given the sentence  $X_t$ , we use the BERT [5] to obtain its textual embedding from the [CLS] token  $E^t \in \mathbb{R}^{m \times d}$  by  $E^t = \Phi_t(X_t)$ , where  $\Phi_t$  is the BERT encoder and  $m \times d$  is the dimensions of the textual embedding. To mitigate intra-modal spurious correlations in the text data, we design a causal intervention network with learnable masks  $M^t \in \mathbb{R}^{m \times d}$ . The values of  $M^t$  range between (0, 1) and are used to filter textual features by training the corresponding parameters.  $\odot$  means element-wise multiplication and  $M^t$  assigns weights to each feature to implement the do-operator.. Through the causal intervention network, we can derive the counterfactual textual features  $\bar{E}^t$  as follows:

$$M^t = \text{Sigmoid}(\text{MLP}(E^t)), \quad (1)$$

$$\bar{E}^t = \text{MLP}(M^t \odot E^t + \text{MLP}(M^t \odot E^t)), \quad (2)$$

where MLP is the multi-layer perception.

The  $\bar{E}^t$  are designed to predict the correct sentiment label accurately. Thus, we use the Cross-Entropy loss function  $L_{ce}$  to guide the model to learn label-relevant textual features as follows,

$$L_t = L_{ce}(\bar{y}^t, y^t), \quad (3)$$

where  $\bar{y}^t = f_t(\bar{E}^t)$  is the prediction of the counterfactual textual features,  $f_t$  represents the textual classifier, and  $y^t$  denotes the textual ground truth.

**Visual Feature Selection.** The image information may interfere with predicting the target sentiment due to inter-modal spurious correlations. To address this issue, we also design a causal intervention network to capture relevant visual features with targets' labels. Given the image  $X_v \in \mathbb{R}^{c \times h \times w}$  with the dimensions of  $c \times h \times w$ , we use the CLIP [31] image encoder  $\Phi_v$  to obtain the image embedding  $E^v = \Phi_v(X_v) \in \mathbb{R}^{m \times d}$ . The causal intervention network for selecting relevant visual features by training learnable masks  $M^v \in \mathbb{R}^{m \times d}$ , which has the same structure as the textual causal intervention network. The counterfactual visual features  $\bar{E}^v$  can be obtained using Eq. (5),

$$M^v = \text{Sigmoid}(\text{MLP}(E^v)), \quad (4)$$

$$\bar{E}^v = \text{MLP}(M^v \odot E^v + \text{MLP}(M^v \odot E^v)). \quad (5)$$

To better align images with their corresponding sentiment labels of targets, we formulate the task as a multi-class classification problem. In practice, the sentiment labels can be marked by a triple (*negative*, *neutral*, *positive*). For instance, if the text associated with an image contains both “*neutral*” and “*positive*” sentiments, its label is defined as (0, 1, 1). We use Binary Cross-Entropy Loss  $L_{bce}$  to constrain predictions of

counterfactual visual features. Therefore, the training loss of visual features is formulated by Eq. (6),

$$L_v = L_{bce}(\bar{y}^v, y^v), \quad (6)$$

where  $\bar{y}^v = f_v(\bar{E}^v)$  denotes the prediction of counterfactual visual features,  $f_v$  represents the image classifier, and  $y^v$  denotes the image ground truth.

### 3.3 Causal Cross-modal Scheduler

An intuitive way for joint learning of selected multi-modal features is to train the sum of  $L_t$  and  $L_v$ . However, this can be sub-optimal as it ignores the fact that varying degrees of spurious correlations in the multi-modal data impact model training differently. To enhance multi-modal learning, we propose a *Causal cross-Modal Scheduler* (CMS) that adapts training based on the perceived causal effects of selected features, assigning greater learning weights to those with stronger causal effects. Inspired by the granger-causal objective [33, 32] for mining causal relationships (illustrations are provided in the Appendix A.1), we propose a novel counterfactual reasoning method to measure the causal effect by computing the loss difference between the counterfactual features and the original features.

**Causal Effect of Textual Feature.** Specifically, the causal effect  $\Delta\epsilon(\bar{E}^t)$  is computed by comparing the losses of  $\bar{E}^t$  and  $E^t$ , which is formulated by Eq. (7),

$$\Delta\epsilon(\bar{E}^t) = \exp(L_{ce}(y^t, \hat{y}^t) - L_{ce}(y^t, \bar{y}^t)), \quad (7)$$

where  $\hat{y}^t = f_t(E^t)$  denotes the prediction of original textual features,  $\bar{y}^t$  is the prediction of the counterfactual textual features, and  $y^t$  denotes the textual ground truth.

**Causal Effect of Visual Feature.** Given the counterfactual visual features  $\bar{E}^v$  and the original visual features  $E^v$ , we also introduce counterfactual reasoning to measure the causal effect of selected visual features. It is formulated by Eq. (8),

$$\Delta\epsilon(\bar{E}^v) = \exp(L_{bce}(y^v, \hat{y}^v) - L_{bce}(y^v, \bar{y}^v)), \quad (8)$$

where  $\hat{y}^v = f_v(E^v)$  represents the prediction of the original visual features,  $\bar{y}^v$  denotes the prediction of counterfactual visual features, and  $y^v$  is the image ground truth.

When  $\Delta\epsilon(\bar{E}^t)$  or  $\Delta\epsilon(\bar{E}^v)$  is greater than 1, it indicates that  $\bar{E}^t$  or  $\bar{E}^v$  generates smaller task loss compared to  $E^t$  or  $E^v$ , suggesting the selected features positively impact sentiment prediction. Conversely, if  $\Delta\epsilon(\bar{E}^t)$  or  $\Delta\epsilon(\bar{E}^v)$  is less than 1, the selected features negatively affect the task. As the model trains, the causal effect is expected to increase or remain stable, allowing the model to learn more relevant features aligned with the target.

**Casual Scheduling Objective.** The larger the causal effect of features, the greater their causal contribution to predicting correct sentiment labels. Based on the causal effects, we then design learnable scheduling weights  $\varphi_{t,b} = \sigma(\alpha \cdot \Delta\epsilon(\bar{E}^t))$  and  $\varphi_{v,b} = \sigma(\beta \cdot \Delta\epsilon(\bar{E}^v))$  (detailed explanation is provided in the Appendix A.2), where  $b$  is the batch index,  $\sigma$  is the softplus function,  $\alpha$  and  $\beta$  are the learnable parameters.  $\varphi_t$  and  $\varphi_v$

aim to perceive the causal effects of the extracted textual and visual features, respectively. Then, we use these learnable scheduling weights to perform a weighted summation of the multi-modal joint training losses to obtain the final scheduled objective, as follows,

$$L = \sum_{b \in D_{train}} (\varphi_{t,b} \cdot L_t + \varphi_{v,b} \cdot L_v), \quad (9)$$

where  $D_{train}$  is the training dataset.

### 3.4 Solving Scheduler via Bi-level Optimization

The scheduler is to optimize the learnable parameters set  $\phi = \{\alpha, \beta\}$  to minimize Eq. (9). Considering the high computational cost of searching for optimal scheduling parameters for multi-modal features in each batch, this approach demands substantial resources. To address this, we introduce a small developing dataset  $D_{dev} = \{(x_b^{dev}, y_b^{dev})\}_b^B$ , which is a small subset sampled from the validation set  $D_v$  [3]. We utilize the objective loss on  $D_{dev}$  to optimize the parameters  $\phi$  to achieve the optimal scheduling weights for multi-modal loss on  $D_{dev}$ . Given the  $\phi$  and MFS model parameter  $\theta$ , our problem can be formulated as a bi-level optimization problem shown as Eq. (11),

$$L_{dev}(\theta^*(\phi)) = \sum_{b \in D_{dev}} (\varphi_{t,b} \cdot L_t + \varphi_{v,b} \cdot L_v), \quad (10)$$

$$\begin{aligned} \phi^* &= \arg \min_{\phi} L_{dev}(\theta^*(\phi)), \\ s.t. \theta^* &= \arg \min_{\theta} L(\theta, \phi), \end{aligned} \quad (11)$$

where  $L_{dev}(\theta^*(\phi))$  is the scheduled training loss in Eq. (9) on  $D_{dev}$ . It is noted that  $\varphi_{t,b}$  and  $\varphi_{v,b}$  are parameterized by  $\phi$ .

In the lower-level optimization, we update the MFS parameter  $\theta$  with the fixed parameter  $\phi$ .  $\theta$  is updated by using the weighted gradient sum of the different modalities as follows,

$$\nabla_{\theta} L(\theta, \phi) = \sum_{b \in D_{train}} (\varphi_{t,b} \cdot \nabla_{\theta} L_t + \varphi_{v,b} \cdot \nabla_{\theta} L_v). \quad (12)$$

In the upper-level optimization, it is expected to compute the gradient  $L_{dev}(\theta^*(\phi))$  to  $\phi$ . Given the indirect dependency of  $L_{dev}(\theta^*(\phi))$  on  $\phi$  through  $\theta$ , we use implicit differentiation to obtain this implicit gradient [22]. Inspired by the *Cauchy-based Implicit Function Theorem* [22], we can leverage the chain rule to systematically derive the gradient of  $L_{dev}(\theta^*(\phi))$  to  $\phi$ ,

$$\begin{aligned} \nabla_{\phi} L_{dev}(\theta^*(\phi)) &= \nabla_{\theta} L_{dev} \cdot \nabla_{\phi} \theta^* \\ &= -\nabla_{\theta} L_{dev} \cdot (\nabla_{\theta}^2 L)^{-1} \cdot \nabla_{\phi} \nabla_{\theta} L|_{(\phi, \theta^*(\phi))}. \end{aligned} \quad (13)$$

The detailed derivation of the implicit gradient is in Appendix A.3. However, directly computing the inverse of the Hessian matrix for deep neural models is often computationally intractable due to its immense size and complexity. To address this, we employ the *K-truncated Neumann series* [3] to approximate this inverse, i.e.,  $(\nabla_{\theta}^2 L)^{-1} \approx$



**Algorithm 1** CMCS Algorithm

---

```

1: Input: datasets:  $D_{train}, D_{dev}$ ; hyperparameters:  $H, K, \eta_1, \eta_2$ 
2: Initialization:  $\theta, \phi, \theta_{opt}$ 
3: while not converge do
4:   // lower-level optimization (update  $\theta$  with fixed  $\phi$ )
5:   for  $t = 0$  to  $H - 1$  do
6:     Calculate the causal effect of textual features by Eq. (7)
7:     Calculate the causal effect of visual features by Eq. (8)
8:     Update model parameters  $\theta$  by
        $\theta = \theta - \eta_1 \nabla_{\theta} L(\theta, \phi)$ 
9:   end for
10:  // upper-level optimization (update  $\phi$  with current  $\theta$ )
11:  Obtain the  $L_{dev}(\theta^*(\phi))$  on  $D_{dev}$  by Eq.(14)
12:  Update scheduling parameters  $\phi$  by
     $\phi = \phi - \eta_2 \nabla_{\phi} L_{dev}$ 
13: end while
14: Return  $\theta_{opt}$ 

```

---

$\sum_{j=0}^K (I - \nabla_{\theta}^2 L)^j$  where  $I$  is the identity matrix. Thus, the implicit gradient  $\nabla_{\phi} L_{dev}(\theta^*(\phi))$  can be calculated as in Eq. (14).

$$\nabla_{\phi} L_{dev} = -\nabla_{\theta} L_{dev} \cdot \sum_{j=0}^K (I - \nabla_{\theta}^2 L)^j \cdot \nabla_{\phi} \nabla_{\theta} L. \quad (14)$$

Algorithm 1 outlines the comprehensive process for optimizing the MFS ( $\theta$ ) and the CMS ( $\phi$ ) based on their gradients. During the lower-level optimization phase,  $\phi$  remains fixed while  $\theta$  is updated using the gradient specified in Eq. (12) at a learning rate of  $\eta_1$ . Rather than aiming for full convergence of  $\theta$ , we employ an efficient H-step optimization approach, inspired by [22], where  $\theta$  undergoes  $H$  iterations of updates before shifting to the upper-level optimization of  $\phi$ . We first evaluate  $L_{dev}$  according to Eq. (14) and then leverage the implicit gradient to update  $\phi$  with a learning rate of  $\eta_2$ . Given  $N$  modalities, the truncated Neumann series number as  $K$ , the time complexity for the gradient backward of CMCS is  $O(N + (K + N)/H)$ . Details are provided in the Appendix A.4.

### 3.5 Model Inference

Cross-modal feature fusion may be suboptimal because images often contain a significant amount of information that may be irrelevant to the sentiment of the target words in the text [52]. To address this, we integrate the prediction scores from both the image and text modalities to make the final sentiment classification during the inference stage. Following previous works [30], we design prompts for images. Specifically, the prompt corresponding to an image is “*it’s a picture of with a target of [label]*”, where  $[label]$  is the target term from the text. For the text prompt  $X_P$ , we use the CLIP [31] text encoder  $\Phi_p$  and prompt embedding  $E^p = \Phi_t(X_P) \in \mathbb{R}^{m \times d}$ . The similarity scores

$\theta = \cos(E^t, E^p)$  can be used to ensemble prediction scores from the two modalities. If the similarity is higher, the image and the text are more relevant, so the proportion of the image will be larger in the final fusion. Finally, the prediction is obtained by computing the weighted sum of the image and text predictions, formulated by Eq.(15),

$$y^h = (1 - \theta) \cdot \bar{y}^t + \theta \cdot \bar{y}^v. \quad (15)$$

## 4 Experiment

### 4.1 Experimental settings

**Dataset.** Following previous works [55, 42], we conduct experiments on two well-known benchmark datasets: Twitter2015 [23] and Twitter2017 [47]. Basic statistics for datasets are summarized in Appendix A.5.

Methods	Twitter2015		Twitter2017	
	Acc	F1	Acc	F1
<b>Text-based</b>				
AE-LSTM*	70.3	63.4	61.7	58.0
MGAN*	71.2	64.2	64.8	61.5
BERT	77.1	71.1	70.2	68.2
<b>Multimodal</b>				
ESAFN <sup>♣</sup>	73.4	67.4	67.8	64.2
TomBERT <sup>♣</sup>	77.2	71.8	70.5	68.0
CapTrBERT <sup>♣</sup>	78.0	73.2	72.3	70.2
JML <sup>♣</sup>	<u>78.7</u>	-	72.7	-
FITE*	78.5	73.9	70.9	68.7
ITM	78.3	<u>74.2</u>	72.6	72.0
CLUE	77.8	72.6	71.7	70.3
GEAR	78.5	73.8	73.1	72.2
VEMP	78.6	74.1	73.0	<u>72.4</u>
AoM	78.3	72.9	<u>73.6</u>	72.0
DQPSA	76.5	-	70.3	-
<b>CMCS(ours)</b>	<b>79.7</b>	<b>75.5</b>	<b>74.3</b>	<b>73.5</b>

Table 1: Results of different methods for TMSC. \* denotes the results from VEMP [42]. <sup>♣</sup> denotes the results from AoM [55].

Due to the space limit, details of baselines are in the Appendix A.6.

### 4.3 Main Results and Analysis

**Performance.** Table 1 presents performance metrics for TMSC. Our proposed CMCS outperforms all text-based models, highlighting the advantages of using multi-modal information. Existing multi-modal sentiment analysis methods often require extensive

**Implementation Details.** Our method is built on BERT [5] and CLIP [31], trained for 30 epochs with a batch size of 32 on TMSC. The learning rate is set to 2e-5, and the hidden sizes of BERT and CLIP are both 1024. All instruction-tuning experiments are conducted using PyTorch on an NVIDIA Tesla V100 GPU.

**Evaluation Metrics.** Following previous studies, we evaluate the performance of our model on the TMSC task by Micro-F1 score (F1) and accuracy (Acc) and report the average of 5 independent training runs as results. To prevent overfitting of the model, the dropout is set to 0.1.

### 4.2 Baselines

We compare our proposed CMCS with the twelve baselines, including AE-LSTM [40], MGAN [7], BERT [5], ESAFN [48], TomBERT [49], CapTrBERT [12], JML [11], FITE [44], ITM [50], CLUE[36], GEAR[35], VEMP[42], AoM [55], and DQPSA [30].

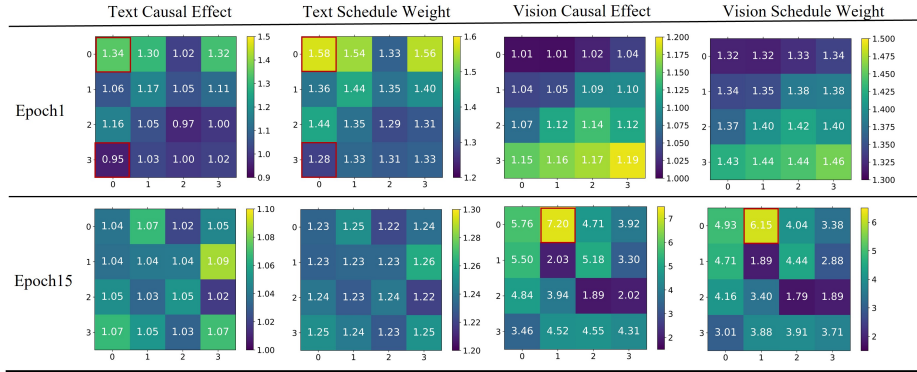


Fig. 4: Visualization of causal effect and scheduling weights.

pre-training on image datasets to align image and text features. In contrast, CMCS does not require such pre-training. Even without pre-training, CMCS exceeds previous multi-modal models in several metrics, achieving 1.0% and 0.9% improvements in accuracy and 1.5% and 0.5% improvements in F1 scores on the Twitter2015 and Twitter2017 datasets, respectively. The accuracy change curves of CMCS compared to the sub-optimal methods in Appendix A.7 show that CMCS achieves a higher accuracy more quickly. The computational efficiency analysis in Appendix A.8 shows that CMCS achieves optimal TMSC performance without significantly increasing training and inference overhead.

#### 4.4 Ablation Study

We study the effectiveness of each component in CMCS, the results are shown in Table 2.

**W/o Visual Modality** shows that after removing the visual modality, the performance declines by 0.6% and 0.9% in accuracy on Twitter2015 and Twitter2017, but 0.4% in F1 on Twitter2015 and 2.4% on Twitter2017. It underscores the importance of the visual modality.

**W/o CMS** indicates that removing the causal cross-modal scheduler results in performance drops of 1.5% and 1.2% in accuracy on Twitter2015 and Twitter2017, respectively. This underscores that measuring the causal effects of different samples helps prioritize those that are easier to learn, ultimately enhancing model performance.

**W/o MFS.** We set the learnable mask values in MFS to a fixed value of 0.5 and tested its ablation results, and it shows that MFS plays a significant role in TMSC.

**W/o (MFS & CMS)** shows that cross-modal causal scheduling is essential; merely extracting features without measuring their effects does not benefit model improvement.

Method	Twitter2015		Twitter2017	
	Acc	F1	Acc	F1
ours	79.7	75.5	74.3	73.5
w/o Visual Modality	79.1	75.1	73.5	71.2
w/o CMS	78.2	73.4	72.4	71.2
w/o MFS	78.3	72.8	72.9	71.6
w/o (MFS & CMS)	78.4	72.5	73.1	71.5

Table 2: Ablation results.

#### 4.5 Further Analysis

To investigate the effectiveness of the causal scheduling, we show the causal effects and scheduling weight of 16 batches at Epoch 1 and Epoch 15, presented in Fig. 4.

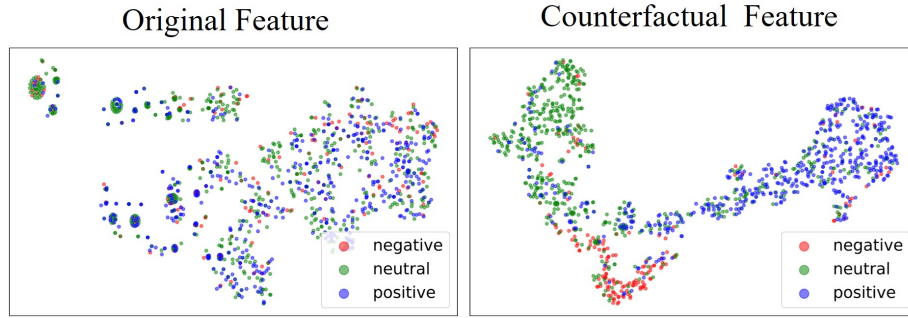


Fig. 5: Textual feature visualization in Twitter2017.

First, within the same epoch, batches with greater causal effects correspond to larger scheduling parameters, demonstrating that our causal cross-modal scheduler effectively adjusts to different samples based on their causal effects. Second, for text features, some causal effects are less than 1 in Epoch 1, while in Epoch 15, almost all exceed 1, suggesting the model becomes more adept at selecting relevant features over time. For image features, a noticeable increase in causal effects with more epochs indicates the presence of irrelevant features in the image data, allowing the model to identify relevant features more easily.

We also use the t-SNE [14] to visualize textual features of test datasets, as shown in Fig. 5 and Fig. 2 (in the Appendix A.7). These two figures illustrate that some original instances are misclassified into other categories. In contrast, CMCS effectively distinguishes the three sentiment polarities through textual causal intervention. This indicates that textual causal intervention can enhance textual sentiment classification.

#### 4.6 Case Study

To intuitively demonstrate the advantage of our method, we compare the predictions of BERT [5], ITM [50], AoM [55], VEMP [42], and CMCS on three test samples, as shown in Fig. 6. In sample (a), BERT misjudges the sentiment by ignoring important words relevant to the target “*Mumias*” while ITM and VEMP errors seem to arise from interference from the visual modality. In sample (b), BERT, VEMP, and ITM incorrectly classify the sentiment of “*Southern NJ*” as negative due to a spurious correlation between the word “Warning” and negative sentiment. In sample (c), both BERT and VEMP misclassify the sentiment of “*Facebook*” as neutral, influenced by surrounding neutral sentiment words. These examples highlight CMCS’s effectiveness in identifying important features related to the target through causal intervention and causal cross-modal scheduling.

## 5 Conclusion

In this paper, we propose a novel *Cross-Modal Causal Scheduling* framework (CMCS) for TMSC, aiming to tackle spurious correlations in multi-modal data. By implementing a *Multi-modal Feature Selection* model (MFS), a *Causal cross-Modal Scheduler*


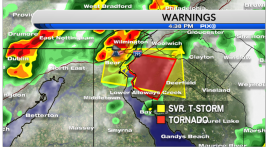

Image			
Sentence	(a) RT @ dailynation : NBK suspends CFO over alleged links to [Mumias] <sub>negative</sub> fiasco	(b) Tornado Warning for Salem County in [Southern NJ] <sub>neutral</sub> until 5pm . @ PIX11News	(c) 17 Awesome # [Facebook] <sub>positive</sub> # Business Page Post Ideas for Small Businesses
BERT	Neutral ×	Negative ×	Neutral ×
ITM	Positive ×	Negative ×	Positive ✓
AoM	Positive ×	Neutral ✓	Positive ✓
VEMP	Negative ✓	Negative ×	Neutral ×
CMCS	Negative ✓	Neutral ✓	Positive ✓

Fig. 6: Comparison of BERT, ITM, AoM, VEMP, and CMCS on three test samples.

(CMS), and the bi-level optimization strategy, CMCS can prioritize relevant multi-modal features. Experimental results on two public datasets validate the effectiveness of our framework in improving sentiment classification.

## Limitations

Since existing datasets for the TMSC task primarily include only the image and text modalities, our method considers information from these two modalities and does not account for other modalities such as audio and video. In the future, we will explore how to effectively utilize additional modalities and investigate multi-modal target sentiment analysis in incremental scenarios.

## Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB3107000.

## References

1. Asgari-Chenaghlu, M., Feizi-Derakhshi, M.R., Farzinvas, L., Balafar, M.A., Motamed, C.: A multimodal deep learning approach for named entity recognition from social media. *Neural Computing and Applications* p. 1905–1922 (Feb 2022)
2. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Jan 2019)
3. Chen, H., Wang, X., Guan, C., Liu, Y., Zhu, W.: Auxiliary learning with joint task and data scheduling. In: *International Conference on Machine Learning, ICML*. pp. 3634–3647 (2022)
4. Chen, Z., Qian, T.: Relation-aware collaborative learning for unified aspect-based sentiment analysis. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jan 2020)

5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* **abs/1810.04805** (2018)
6. Fan, F., Feng, Y., Zhao, D.: Multi-grained attention network for aspect-level sentiment classification. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Jan 2018)
7. Fan, F., Feng, Y., Zhao, D.: Multi-grained attention network for aspect-level sentiment classification. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Jan 2018)
8. Jia, M., Xie, C., Jing, L.: Debiasing multimodal sarcasm detection with contrastive learning. In: *AAAI Conference on Artificial Intelligence* (2023)
9. Jim, J.R., Talukder, M.A.R., Malakar, P., Kabir, M.M., Nur, K., Mridha, M.F.: Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal* p. 100059 (2024)
10. Johansson, F.D., Shalit, U., Sontag, D.: Learning representations for counterfactual inference. In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. pp. 3020–3029 (2016)
11. Ju, X., Zhang, D., Xiao, R., Li, J., Li, S., Zhang, M., Zhou, G.: Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In: *Proceedings of the 2021 conference on empirical methods in natural language processing*. pp. 4395–4405 (2021)
12. Khan, Z., Fu, Y.: Exploiting bert for multimodal target sentiment classification through input space translation. In: *Proceedings of the 29th ACM International Conference on Multimedia* (Oct 2021)
13. Kim, J., Lee, B.K., Ro, Y.M.: Demystifying causal features on adversarial examples and causal inocular for robust network by adversarial instrumental variable regression. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12032–12042 (2023)
14. Laurens, Maaten, V.D., Hinton, Geoffrey: Visualizing data using t-sne. *Journal of Machine Learning Research* **9**(2605), 2579–2605 (2008)
15. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, A., Huang, F.: A tutorial on energy-based learning. *Predicting structured data* (Jan 2006)
16. Li, R., Chen, H., Feng, F., Ma, Z., Wang, X., Hovy, E.: 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)* (Jan 2021)
17. Liang, S., Wei, W., Mao, X.L., Wang, F., He, Z.: Bisyn-gat+: Bi-syntax aware graph attention network for aspect-based sentiment analysis (Apr 2022)
18. Ling, Y., Yu, J., Xia, R.: Vision-language pre-training for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2204.07955* (Apr 2022)
19. Liu, B.: *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2012)
20. Liu, B.: *Sentiment analysis and opinion mining*. Springer Nature (2022)
21. Liu, D., Li, L., Tao, X., Cui, J., Xie, Q.: Descriptive prompt paraphrasing for target-oriented multimodal sentiment classification. In: *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. pp. 4174–4186. Association for Computational Linguistics (2023)
22. Lorraine, J., Vicol, P., Duvenaud, D.: Optimizing millions of hyperparameters by implicit differentiation. In: *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS. Proceedings of Machine Learning Research*, vol. 108, pp. 1540–1552. PMLR (2020)
23. Lu, D., Neves, L., Carvalho, V., Zhang, N., Ji, H.: Visual attention model for name tagging in multimodal social media pp. 1990–1999 (2018)

24. Lv, Y., Wei, F., Cao, L., Peng, S., Niu, J., Yu, S., Wang, C.: Aspect-level sentiment analysis using context and aspect memory network. *Neurocomputing* p. 195–205 (Mar 2021)
25. Moon, S., Neves, L., Carvalho, V.: Multimodal named entity recognition for short social media posts. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (Jan 2018)
26. Oh, S., Lee, D., Whang, T., Park, I., Gaeun, S., Kim, E., Kim, H.: Deep context- and relation-aware learning 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 2: Short Papers)* (Jan 2021)
27. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2**(1-2), 1–135 (2007)
28. Pang, S., Xue, Y., Yan, Z., Huang, W., Feng, J.: Dynamic and multi-channel graph convolutional networks for aspect-based sentiment analysis. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Jan 2021)
29. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Cambridge University Press (2000)
30. Peng, T., Li, Z., Wang, P., Zhang, L., Zhao, H.: A novel energy based model mechanism for multi-modal aspect-based sentiment analysis. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 18869–18878 (2024)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
32. Schwab, P., Karlen, W.: Cxplain: Causal explanations for model interpretation under uncertainty. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*. pp. 10220–10230 (2019)
33. Schwab, P., Miladinovic, D., Karlen, W.: Granger-causal attentive mixtures of experts: Learning important features with neural networks. In: *The Thirty-Third AAAI Conference on Artificial Intelligence*. pp. 4846–4853 (2019)
34. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: Generalization bounds and algorithms. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. pp. 3076–3085 (2017)
35. Sun, T., Ni, J., Wang, W., Jing, L., wei Wei, Y., Nie, L.: General debiasing for multimodal sentiment analysis. *Proceedings of the 31st ACM International Conference on Multimedia* (2023)
36. Sun, T., Wang, W., Jing, L., Cui, Y., Song, X., Nie, L.: Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. *Proceedings of the 30th ACM International Conference on Multimedia* (2022)
37. Wang, Q., Ding, K., Liang, B., Yang, M., Xu, R.: Reducing spurious correlations in aspect-based sentiment analysis with explanation from large language models. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. pp. 2930–2941 (2023)
38. Wang, Q., Ding, K., Liang, B., Yang, M., Xu, R.: Reducing spurious correlations in aspect-based sentiment analysis with explanation from large language models. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. pp. 2930–2941. Association for Computational Linguistics, Singapore (Dec 2023)
39. Wang, X., Cai, J., Jiang, Y., Xie, P., Tu, K., Lu, W.: Named entity and relation extraction with multi-modal retrieval. *Cornell University - arXiv, Cornell University - arXiv* (Dec 2022)
40. Wang, Y., Huang, M., zhu, x., Zhao, L.: Attention-based lstm for aspect-level sentiment classification. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Jan 2016)

41. Xu, N., Mao, W., Chen, G.: Multi-interactive memory network for aspect based multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* p. 371–378 (Sep 2019)
42. Yang, B., Li, J.: Visual elements mining as prompts for instruction learning for target-oriented multimodal sentiment classification. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6062–6075 (2023)
43. Yang, D., Li, M., Xiao, D., Liu, Y., Yang, K., Chen, Z., Wang, Y., Zhai, P., Li, K., Zhang, L.: Towards multimodal sentiment analysis debiasing via bias purification. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision – ECCV 2024*, pp. 464–481. Springer Nature Switzerland, Cham (2025)
44. Yang, H., Zhao, Y., Qin, B.: Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In: *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 3324–3335 (2022)
45. Yang, M., Wang, Z., Xu, Q., Li, C., Xu, R.: Leveraging hierarchical semantic-emotional memory in emotional conversation generation. *CAAI Transactions on Intelligence Technology* **8**(3), 824–835 (2023)
46. Yang, X., Feng, F., Ji, W., Wang, M., seng Chua, T.: Deconfounded video moment retrieval with causal intervention. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021)
47. Yu, J., Jiang, J.: Adapting bert for target-oriented multimodal sentiment classification. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (Aug 2019)
48. Yu, J., Jiang, J., Xia, R.: Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**, 429–439 (2020)
49. Yu, J., Jiang, J., Yang, L., Xia, R.: Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: *Meeting of the Association for Computational Linguistics* (2020)
50. Yu, J., Wang, J., Xia, R., Li, J.: Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In: Raedt, L.D. (ed.) *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 4482–4488. International Joint Conferences on Artificial Intelligence Organization (7 2022), main Track
51. Zhang, Q., Fu, J., Liu, X., Huang, X.: Adaptive co-attention network for named entity recognition in tweets. *Proceedings of the AAAI Conference on Artificial Intelligence* (Nov 2022)
52. Zhao, F., Li, C., Wu, Z., Ouyang, Y., Zhang, J., Dai, X.: M2DF: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9057–9070. Association for Computational Linguistics, Singapore (Dec 2023)
53. Zhao, F., Wu, Z., Long, S., Dai, X., Huang, S., Chen, J.: Learning from adjective-noun pairs: A knowledge-enhanced framework for target-oriented multimodal sentiment classification. In: *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*, Gyeongju, Republic of Korea, October 12–17, 2022, pp. 6784–6794. International Committee on Computational Linguistics (2022)
54. Zhou, J., Lin, Y., Chen, Q., Zhang, Q., Huang, X., He, L.: Causalabsc: Causal inference for aspect debiasing in aspect-based sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **32**, 830–840 (2024)
55. Zhou, R., Guo, W., Liu, X., Yu, S., Zhang, Y., Yuan, X.: Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2306.01004* (2023)