Information Fusion 114 (2025) 102725

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

AtCAF: Attention-based causality-aware fusion network for multimodal sentiment analysis

Changqin Huang [a], Jili Chen [a], Qionghao Huang [a,*], Shijin Wang [b], Yaxin Tu [a], Xiaodi Huang [c]

[a] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, 321004, China
[b] State Key Laboratory of Cognitive Intelligence & iFLYTEK AI Research, iFLYTEK Co., Ltd, Hefei, China
[c] School of Computing, Mathematics and Engineering, Charles Sturt University, Albury, NSW 2640, Australia

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Causal inference
Multimodal fusion

ABSTRACT

Multimodal sentiment analysis (MSA) involves interpreting sentiment using various sensory data modalities. Traditional MSA models often overlook causality between modalities, resulting in spurious correlations and ineffective cross-modal attention. To address these limitations, we propose the **A**ttention-based **C**ausality-**A**ware **F**usion (AtCAF) network from a causal perspective. To capture a causality-aware representation of text, we introduce the **C**ausality-**A**ware **T**ext **D**ebiasing **M**odule (CATDM) utilizing the front-door adjustment. Furthermore, we employ the **C**ounterfactual **Cro**ss-modal **A**ttention (CCoAt) module integrate causal information in modal fusion, thereby enhancing the quality of aggregation by incorporating more causality-aware cues. AtCAF achieves state-of-the-art performance across three datasets, demonstrating significant improvements in both standard and Out-Of-Distribution (OOD) settings. Specifically, AtCAF outperforms existing models with a 1.5% improvement in ACC-2 on the CMU-MOSI dataset, a 0.95% increase in ACC-7 on the CMU-MOSEI dataset under normal conditions, and a 1.47% enhancement under OOD conditions. CATDM improves category cohesion in feature space, while CCoAt accurately classifies ambiguous samples through context filtering. Overall, AtCAF offers a robust solution for social media sentiment analysis, delivering reliable insights by effectively addressing data imbalance. The code is available at https://github.com/TheShy-Dream/AtCAF.

1. Introduction

Sentiment is crucial to human interaction, shaping communication and decisions [1]. As social media and sensor technologies evolve, multimodal sentiment analysis harnesses diverse data like text, audio, and video to accurately gauge sentiment scores [2]. Prior research in multimodal sentiment analysis has primarily concentrated on facilitating interaction and integration among modalities. Some researchers use tensor-based methods to obtain modal interaction representations [3–5]. In addition, several studies use attentional mechanisms for cross-modal modeling [6–10]. Some researchers have also designed auxiliary tasks and self-supervised modules to help reduce the gap between modalities [11–13]. In essence, most of these methodologies are devised to enhance the extraction of consistent information across modalities and the reduction of redundant information by either introducing novel model architectures or tasks.

Although traditional Multimodal Sentiment Analysis (MSA) models have shown enhancements in accuracy, they typically assess modality similarity based on their co-occurrence when labels are provided for improved modal fusion. However, this fusion approach, reliant on co-occurrence and statistical correlations, is suboptimal as it fails to capture the causal relationship underlying modality interaction accurately and cannot offer causal reasoning for prediction outcomes. Consequently, two primary issues persist, leading traditional baseline models to make erroneous predictions, as depicted in Fig. 1: dataset bias and confusion in multimodal fusion. Due to data bias, the baseline model is susceptible to the influence of imbalanced category distributions, leading to a spurious correlation between text and labels. Specifically, upon analyzing the distribution of specific tokens in the samples depicted in Fig. 1 on the training set using the BERT tokenizer, we observe that words such as “movie”, “umm”, and even the letter “t” predominantly appear in the negative category. Consequently, the model may inadvertently learn from this distribution that the presence of the word “movie” is indicative of negative sentiment—an inherently absurd conclusion. As we understand, the word “movie” does not inherently contain any sentimental cues and thus cannot reliably serve as a factor for sentiment judgment. Such spurious correlations [14–17] significantly impact the accuracy of multimodal sentiment analysis and markedly diminish the reliability

* Corresponding author.
E-mail address: qhhuang@m.scnu.edu.cn (Q. Huang).

https://doi.org/10.1016/j.inffus.2024.102725
Received 10 May 2024; Received in revised form 7 September 2024; Accepted 29 September 2024
Available online 2 October 2024

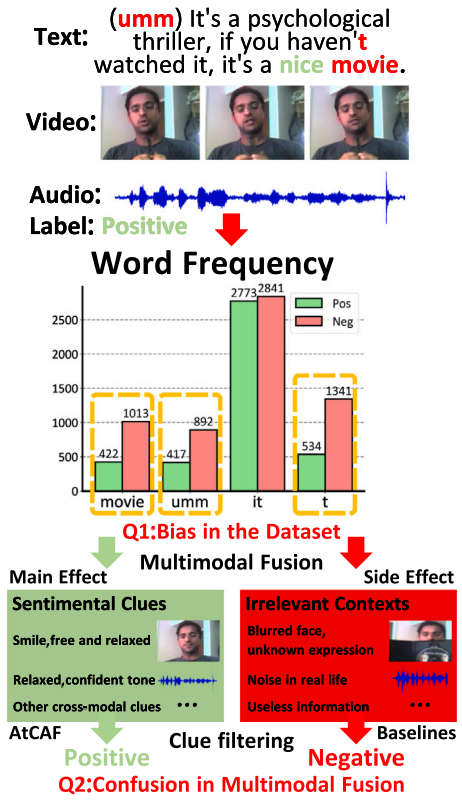


Fig. 1. Two primary issues leading to incorrect predictions by baseline models on test samples (indicated by red arrows) are dataset bias and confusion in multimodal fusion. Dataset bias is evident in the spurious correlations between text and labels, as depicted in the bar plot. Confusion in multimodal fusion arises from the challenge of discerning between causal cues (Main effect) and irrelevant context (Side effect). Our AtCAF model adeptly distinguishes causal clues and facilitates accurate predictions (illustrated by green arrows).

and robustness of multimodal sentiment analysis models in real-world applications. Therefore, a fundamental question arises: ***(Q1): How can we obtain the causality-aware representation of the text modality before modal fusion?*** Multimodal fusion plays a crucial role in mitigating such spurious correlations. For instance, in the text depicted in Fig. 1, which contains both "movie" and "nice", the model tends to prioritize "nice" when interacting with positive facial expressions and tone features, disregarding the spurious correlation between "movie" and the label. Nevertheless, exploring causality in multimodal fusion remains limited, leading to ambiguity: What types of information should be integrated? While humans inherently navigate a multimodal world, capable of understanding the contextual nuances and core content to make unbiased judgments, machines often operate within a likelihood-based framework, grappling with the differentiation between main effects and side effects. Thus, another critical challenge arises: ***(Q2): In which way can we apply causality-aware multimodal fusion for robust sentimental inference?***

To address both the aforementioned questions, we propose the **At**ention-based **C**ausality-**A**ware **F**usion network (AtCAF). We integrate more causality-aware information to overcome the impact of spurious correlations on prediction outcomes. The overall workflow is illustrated in Fig. 2. Specifically, to answer ***(Q1)***, we first employ a structural causal model (SCM) [18] to describe the causal diagram of multimodal sentiment analysis. For the text modality, we design the **C**ausality-**A**ware **T**ext **D**ebiasing **M**odule (CATDM). Inspired by the front-door adjustment [18], this module mitigates the influence of confounders by consolidating information from the global dictionary, which encompasses global textual features. To answer ***(Q2)***, we reanalyze the cross-modal attention mechanism in a causal graph and introduce a novel **C**ounterfactual **Cro**ss-modal **At**tention (CCoAt) module, which provides counterfactual reasoning for conventional cross-modal attention. Leveraging counterfactual theory, this module effectively filters out irrelevant contexts during modal fusion. Moreover, the CCoAt module demands minimal computational resources during training and imposes no additional computational costs during inference. Furthermore, our experiments on CMU-MOSI [19], CMU-MOSEI [20], UR-FUNNY [21] and additional OOD tests showcase that our proposed AtCAF significantly enhances sentiment prediction accuracy, attaining state-of-the-art performance. Ablation studies affirm the effectiveness of each component within our experimental network for sentiment predictions. The contributions of our work can be summarized as follows:

- We propose the novel Attention-based causality-aware fusion network (AtCAF) for multimodal sentiment analysis, which captures causal relationships in the training data to construct a comprehensive causality chain that effectively traces the causal trajectory from user inputs to model outputs.
- Building upon the causal diagram depiction of multimodal sentiment analysis, we extend the front-door adjustment to multimodal learning and introduce the Causality-Aware Text Debiasing Module (CATDM) to acquire a causality-aware representation of the text modality.
- Employing counterfactual theory, we design the Counterfactual Cross-modal Attention (CCoAt) module. To the best of our knowledge, this is the first study to discover causal relationships in multimodal fusion in deep learning-based multimodal sentiment analysis.
- Extensive experiments conducted on CMU-MOSI, CMU-MOSEI, UR-FUNNY datasets, and additional OOD tests on CMU-MOSEI affirm the effectiveness and generalization capability of our proposed AtCAF.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 outlines the methodology of our proposed AtCAF. Section 4 reports experimental results and analysis. Section 5 concludes the entire paper and describes future work.

2. Related work

This section reviews related work on multimodal sentiment analysis and causal inference.

2.1. Multimodal sentiment analysis

Multimodal sentiment analysis entails interpreting emotions expressed through diverse channels such as text, audio, and visuals. Its objective is to comprehend and categorize sentiments by amalgamating cues from multiple modalities, thereby facilitating a comprehensive understanding of sentimental states for applications in human–computer interaction and affective computing.

Previous work has focused on modal fusion methods to capture modal coherence information. Zedeh et al. [3] use a tensor fusion layer to explicitly aggregate unimodal, bimodal, and trimodal interactions for sentiment prediction. Zedeh et al. [22] design a memory fusion network (MFN) to capture intra-view dynamics and inter-view interactions over time. Hazarika et al. [23] propose a novel framework that learns the modality-invariant and modality-specific representations to capture the distinctive features of each modality and complement the invariant representations. Transformer [24] addresses the challenge of long-range dependencies in modal representations. Tsai et al. [6] extend traditional cross-attention to cross-modal attention to automatically align and fuse multimodal information. The Multimodal Adaptation Gate (MAG) is introduced by Rahman et al. [25] to enable the acceptance

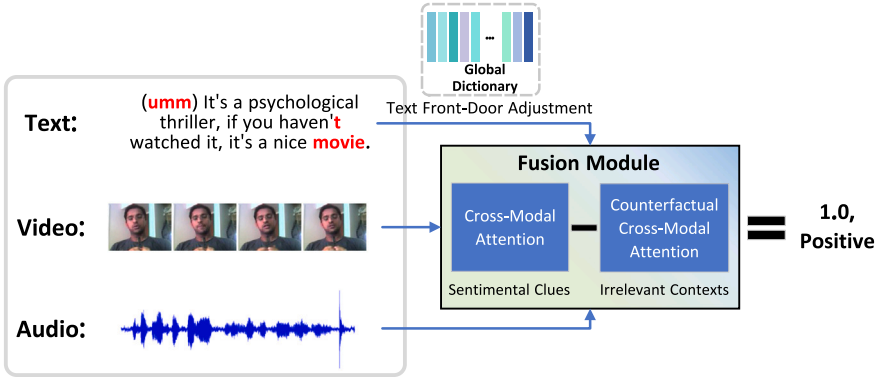


Fig. 2. The workflow of our proposed AtCAF for the test sample.

and integration of nonverbal information during fine-tuning by pre-trained BERT and XLNET. Sun et al. [26] represent interaction features using the outer product of text, audio, and video. Han et al. [12] leverage parametric and non-parametric methods to maximize mutual information in multimodal fusion for sentiment analysis. Kim et al. [13] employ two pre-training tasks, namely Multimodal Masked Language Modeling and Alignment Prediction, to facilitate information fusion across modalities in the BERT encoder.

The research above has indeed advanced accuracy in multimodal sentiment analysis. However, it overlooks the presence of spurious correlations between modalities. Consequently, the knowledge acquired from the training set may not be transferable or reliable when applied to the test set or real-world scenarios, leading to inaccuracies and unreliability. In response to this challenge, we propose a novel front-door adjustment method tailored for multimodal learning, which conducts intra-sample and inter-sample sampling for the endogenous textual modality, while keeping the exogenous visual and audio modalities unchanged to mitigate potential text bias in multimodal emotion analysis. To the best of our knowledge, AtCAF is the first work to extend front-door adjustment to alleviate textual spurious correlations in multimodal sentiment analysis.

2.2. Causal inference

Indeed, numerous studies have highlighted the presence of spurious correlations in deep learning [27–29]. Causal inference, extensively employed in deep learning, eliminates spurious correlations within complex datasets. By discerning authentic causal relationships, causal inference enriches models, fostering robustness in analysis and prediction.

Researchers must select various causal inference tools to address different forms of debiasing. Wang et al. [14] mitigate confounders in object detection using the back-door criterion and the do-operator. Rao et al. [30] introduce counterfactual attention learning (CAL) to enhance the learning of robust features within traditional attention mechanisms for visual categorization and re-identification. Yang et al. [16] employ the front-door criterion to construct causal attention in visual-language models. Liu et al. [31] employ the front-door and back-door criteria to eliminate spurious correlations in textual and visual modalities for vision question answering (VQA). Huang et al. [32] devise a counterfactual attention generator (CAG) to automatically guide the factual attention module in learning invariant features and making sharp predictions for facial expression recognition (FER). Causal inference is also applied to mitigate spurious correlations by establishing training processes from a causal perspective rather than the traditional statistical viewpoint. Sun et al. [33] identify spurious correlations within textual data in multimodal sentiment analysis and effectively mitigated them using a counterfactual framework. Moreover, Sun et al. [34] employ a novel GMAE (Generalized Mean Absolute Error) loss function to disentangle robust and biased features in each modality, aiming to reduce spurious correlations between features and sentiment labels.

However, most studies focus on bias reduction within unimodal domains, overlooking the aspect of cross-modal causal discovery. Hence, our Attention-based Causality-Aware Fusion (AtCAF) network addresses the mitigation of spurious correlations in the text modality as an integral part of the representation learning process. Subsequently, we utilize counterfactual cross-modal attention to facilitate cross-modal causal discovery. We propose an auxiliary optimization module with cross-attention by estimating causal effects under different modalities. To the best of our knowledge, the work of decoupling cross-attention and then performing counterfactual filtering for cross-modal causal discovery is unprecedented.

3. AtCAF

In this section, we define the multimodal sentiment analysis task, present preliminary knowledge, and formulate causal assumptions. Subsequently, we provide a detailed introduction to our proposed network.

3.1. Task definition

We use $D = \{T_i, V_i, A_i, Y_i\}_1^N$ to denote the set of training samples. Each sample is a video clip containing three modalities: text modality T_i, video modality V_i, audio modality A_i, and a sentimental label Y_i. The multimodal sentiment analysis task aims to construct a unified model F_θ with learnable parameters θ that processes all three modalities together and outputs a single sentiment analysis score $\hat{Y}_i$ ranging from −3 to 3 or sentimental logits $\hat{Y}_i$ (determined by the datasets) as follows:

$$\hat{Y}_i = F_\theta(T_i, V_i, A_i). \tag{1}$$

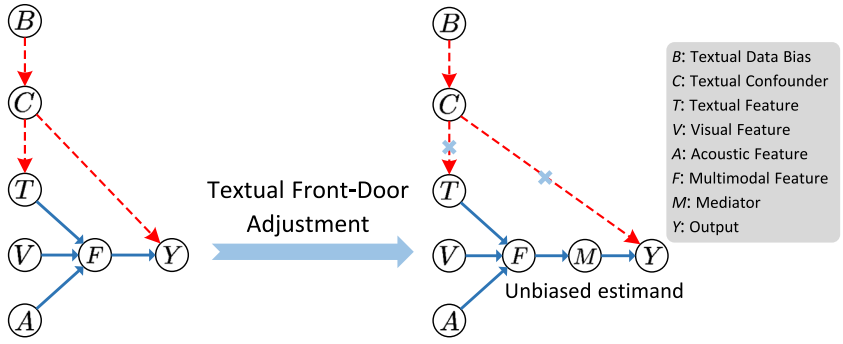
Table 1 lists the basic notations used in this paper.

3.2. Preliminary and causal relationships

In this section, we refer to the Structural Causal Model (SCM) [18] to present the preliminary knowledge and the causal relationships addressed in this paper. The subsequent content of this paper is built upon these causal assumptions.

Causal Graph. The causal graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is a graphical representation that depicts the causal relationships among a set of variables or factors. It consists of nodes $\mathcal{V}$, which represent the variables, and directed edges $\mathcal{E}$, which indicate the direction of causation between variables [35].

The Causal Graph of Multimodal Sentiment Analysis. Based on the traditional Multimodal Sentiment Analysis (MSA) procedure, we illustrate the causal relationships in the MSA task in Fig. 3(a),



(a) Original MSA causal graph. (b) Deconfounded MSA causal graph.

Fig. 3. The proposed textual front-door adjustment. The blue line represents the true causal relationship in MSA, while the red dashed line represents the bias introduced by confounders.

i.e. the causal relationships among the bias B, confounder C in text $\{T_i\}_1^N$, audio $\{A_i\}_1^N$, video $\{V_i\}_1^N$, multimodal fusion feature F and model output Y. The confounder C denotes the frequent occurrence of specific concepts, and text bias B refers to the strong correlation between concepts and labels, despite the absence of genuine causal relationships. The presence of text bias B in the training set results in the emergence of the confounder C, which subsequently impacts multimodal fusion and model output. Specifically, we use two causal paths to describe the causal relationships from the input $\{T_i, V_i, A_i\}$ to the output Y_i: **Path 1:** $T \leftarrow C \rightarrow Y$ is a back-door path. This path reflects the spurious correlation between text and labels in the training set. Suppose the concept of "movie" frequently co-occurs with a negative sentiment label in the training set. In that case, the model will correlate these two concepts. Therefore, when "movie" appears in the testing set, the model will infer a negative sentiment score without considering other sentimental clues. **Path 2:** $\{T, V, A\} \rightarrow F \rightarrow Y$ describes the fusion of modalities to generate the output. The model will perform reliable knowledge extraction and multimodal interaction through this path to obtain the multimodal representation F. For example, when "movie" and "nice" simultaneously appear in the text and interact with happy facial expressions and excited tones, the model will filter out the spurious correlation related to "movie" and focus on the genuine clues related to "nice". As such, this path represents a real causal effect.

In summary, traditional Multimodal Sentiment Analysis (MSA) models predominantly emphasize modal fusion, neglecting the influence of spurious correlation induced by the back-door path. Consequently, the sentimental cues captured in the training set may not generalize effectively to the testing set [16].

3.3. Overall architecture

In this section, we provide a detailed description of our proposed AtCAF. Initially, we conduct unimodal feature extraction to acquire unimodal temporal features. For biased textual modalities, we utilize front-door adjustment to mitigate spurious correlations with a global dictionary and introduce the Causality-Aware Text Debiasing Module (CATDM). Lastly, we devise Counterfactual Cross-modal Attention (CCoAt) to enhance the original cross-modal attention mechanism for capturing causality-aware sentimental clues. The overall architecture diagram is illustrated in Fig. 4.

3.3.1. Unimodal feature extraction

To transform the multimedia input into a tensor representation for input into the model, we utilize COVERAP [36] to extract audio features and Facet [37] to extract video features as described in 4.4. Considering the diverse characteristics of different modalities, we employ distinct modules to extract modal features. Specifically, we first pad the text, video, and audio in batches independently and utilize Long Short-Term Memory (LSTM) networks [38] to extract temporal features from each modality separately, while employing Conv1D for remapping purposes in multimodal fusion [6]. As for the text modality, we utilize a pre-trained transformer as the backbone network. In this process, features from all modalities are extracted independently for modal fusion:

$$\begin{aligned} h_t &= Transformer(I_t), \\ F_m &= Conv1d(LSTM(I_m)), m \in \{a, v\}, \end{aligned} \tag{2}$$

where $F_m, m \in \{a, v\}$, denotes the audio and visual features after unimodal feature extraction. The text modality is h_t after passing through the text encoder.

Table 1
Basic notation reference.

Notation	Description
t	The text modality
a	The audio modality
v	The video modality
m	A specific modality
I_m	The input of modality m
F_m	The representation of m after feature extraction
l_m	The sequence length of m
d_m	The feature dimension of m
$Transformer(\cdot)$	The pre-trained text encoder
$LSTM(\cdot)$	The LSTM network
$Conv1d(\cdot)$	The one-dimensional convolution
$do(\cdot)$	The do operator
$\mathbb{E}(\cdot)$	The calculation of the expectation
$Pool(\cdot)$	The mean pooling operation
$MHA(\cdot)$	The multi-head attention operation
$LN(\cdot)$	The layer normalization operation
$FFN(\cdot)$	The feedforward network
$softmax(\cdot)$	The softmax layer
$[\cdots;\cdots]$	The concatenation operation
$MLP(\cdot)$	The multilayer perceptron

3.3.2. Causality-aware text debiasing module

As mentioned in Section 3.2, traditional MSA models suffer from spurious correlation introduced by text. To address the issue of spurious correlation in the text and obtain the causality-aware representation of the text modality (**referred to as Q1 in the Introduction**), we leverage causal intervention for further analysis [39].

Theoretical Analysis of the Debiasing Method. For the text modality, the absence of the sampling process renders it impossible to observe the text confounder C directly. Fortunately, the front-door adjustment presents a novel causal pathway for unobserved confounders. This adjustment technique allows for the manipulation of contexts, enabling the evaluation of the confidence of the knowledge learned across different samples. For instance, one can isolate the occurrence of "movie" associated with a negative label from the sentence and place it separately within the context of other samples to assess its causal

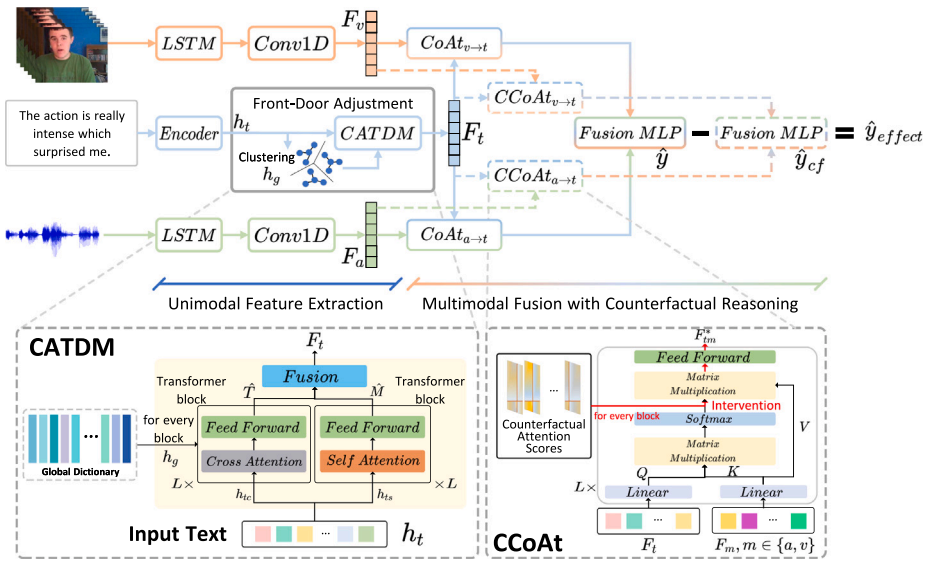


Fig. 4. The overall architecture of our proposed AtCAF. We represent the unimodal feature extraction process with solid-colored lines, depict the multimodal fusion process with gradient-colored lines, and illustrate the incorporation of counterfactual information in the fusion process using gradient dashed lines. The lower left subplot depicts the process of CATDM. The text h_t is divided into two branches. One branch uses self-attention for in-sample learning, while the other branch interacts with the global dictionary constructed through clustering using cross-attention to obtain a causality-aware text representation F_t. The lower right subplot illustrates the workflow of a counterfactual cross-modal attention block. Building upon the traditional cross-modal attention, we sample some attention scores from a predefined distribution to replace the original cross-attention scores. We have marked the implementation process of the counterfactual cross-modal intervention with red solid lines.

effect on the label. If "movie" exhibits an insignificant causal effect on the label in other contexts, it will be reassigned a relatively more minor token-wise weight for inference within the front-door adjustment procedure.

Specifically, to apply the front-door adjustment [18], we need to introduce a mediator M [31] to construct a new causal pathway $T \to F \to M \to Y$ to transmit the causal effect as illustrated in Fig. 3(b). Based on causal graphs, we reformulate the MSA task as follows:

$$P(Y|T,V,A) = \sum_k P(M=k|T,V,A)P(Y|M=k), \tag{3}$$

where k denotes the selection of knowledge from T based on audio A and video V to make the final prediction Y. Therefore, the MSA task is divided into the following two parts. $\{T,V,A\} \to F \to M$ denotes fusion module with multimodal knowledge extractor. $M \to Y$ denotes sentiment predictor. To block the back-door path $T \leftarrow C \to Y$, we introduce the do-operator $do(\cdot)$ for causal intervention [18,31]. For example, $do(A = \bar{A})$ signifies forcibly setting the variable A to $\bar{A}$, thereby severing all causal connections from its parental nodes in the causal graph. In other words, we cut off the causal path $T \leftarrow C$. Thus, the intervention probability can be represented as:

$$P(Y|do(T),V,A) = \sum_k P(M=k|do(T),V,A)P(Y|do(M=k)), \tag{4}$$

there is no unblocked back-door path between T and M, so we can get:

$$P(M=k|do(T),V,A) = P(M=k|T,V,A), \tag{5}$$

however, an unblocked back-door path between M and Y exists. To address this, we apply the back-door adjustment:

$$P(Y|do(M=k)) = \sum_t P(T=t)P(Y|T=t,M=k). \tag{6}$$

By combining Eqs. (4), (5), and (6), we can derive the formula for the unbiased estimand from T,V,A to Y:

$$P(Y|do(T),V,A) = \sum_k P(M=k|T,V,A)\sum_t P(T=t)P(Y|T=t,M=k), \tag{7}$$

where the direct calculation of the above conditional probabilities involves a large amount of iterative computation and global statistics, so we use deep neural networks to handle the above probability estimation [16].

Implementation with the deep neural network of CATDM. To implement Eq. (7) through a deep learning framework, the most straightforward idea is to parameterize a module z with the softmax function to estimate the probabilities [40]:

$$\begin{aligned} &P(Y|do(T),V,A) = \\ &\sum_k P(M=k|T,V,A)\sum_t P(T=t)P(Y|T=t,M=k) \\ &= \mathbb{E}_{[M|T,V,A]}\mathbb{E}_{[T']}[p(Y|M,T')] \\ &= \underbrace{\mathbb{E}_{[M|T,V,A]}}_{ISKS}\underbrace{\mathbb{E}_{[T']}}_{CSKS}[softmax(z(M,T'))], \end{aligned} \tag{8}$$

where T' denotes the text modality from another sample. We parameterize the module z to estimate the probability value $p(Y|M,T')$ for in-sample knowledge selection (ISKS) and cross-sample knowledge selection (CSKS) [31,41]. ISKS involves extracting pertinent information solely from the current input sample, whereas CSKS entails estimating biases and recalibrating by integrating the current sample with others from the training set. Nevertheless, sampling across the entirety of available samples incurs substantial computational overhead. We use the Normalized Weighted Geometric Mean (NWGM) [42] method to estimate sampling down to the feature level:

$$\begin{aligned} &P(Y|do(T),V,A) = \mathbb{E}_{[M|T,V,A]}\mathbb{E}_{[T']}[softmax(z(M,T'))] \\ &\overset{NWGM}{\approx} softmax(\mathbb{E}_{[M|T,V,A]}\mathbb{E}_{[T']}[z(M,T')]), \end{aligned} \tag{9}$$

furthermore, according to the linearity of expectation and the properties of neural networks, we can put the expectation operation into $z(\cdot)$ and represent it as:

$$\begin{aligned} &P(Y|do(T),V,A) \approx softmax(z(\mathbb{E}_{[M|T,V,A]}[M],\mathbb{E}_{[T']}[T'])) \\ &= softmax(z(\underbrace{\sum_k P(M=k|g_1(T))k}_{ISKS},\underbrace{\sum_t P(T=t|g_2(T))v}_{CSKS})) \\ &= softmax(z(\hat{M},\hat{T})), \end{aligned} \tag{10}$$

where $\hat{M}$ denotes the in-sample knowledge selection process for estimating M, $\hat{T}$ denotes the cross-sample knowledge selection process for estimating T', and g_1 and g_2 are mapping functions, k and t represent the selected knowledge in the corresponding process.

Based on Eq. (10), we build the **C**ausality-**A**ware **T**ext **D**ebiasing **M**odule (CATDM) shown in the lower left subplot of Fig. 4. Specifically, we use the K-means algorithm to initialize a global dictionary $h_g \in \mathbb{R}^{N\times d_t}$ with over the textual modalities of the entire training set T_g [16]:

$$h_g = KMeans(Pool(Transformer(T_g)), N). \tag{11}$$

where N denotes the size of the dictionary. We design a dual-branch network to estimate $\hat{M}$ and $\hat{T}$ separately. One of the branches is used to perform self-attention calculations on the input text tokens h_t to obtain an estimate of the internal knowledge $\hat{M}$ within the sample. The other branch takes the input h_t as the query vector and uses the global dictionary as the key and value vectors for cross-attention to estimate $\hat{T}$ [40]. We employ an L-layer Transformer that leverages self-attention and another L-layer Transformer that utilizes cross-attention to implement the estimations:

$$\begin{aligned} \tilde{M}^i &= MHA(LN(h_{ts}^{i-1})) + LN(h_{ts}^{i-1}), \\ h_{ts}^i &= FFN(LN(\tilde{M}^i)) + LN(\tilde{M}^i), \end{aligned} \tag{12}$$

$$\begin{aligned} \tilde{T}^i &= MHA(LN(h_{tc}^{i-1}), LN(h_g)) + LN(h_{tc}^{i-1}), \\ h_{tc}^i &= FFN(LN(\tilde{T}^i)) + LN(\tilde{T}^i), \end{aligned} \tag{13}$$

where $MHA(X)$ represents self-attention operation on X, while $MHA(X,Y)$ denotes cross-attention operation with X as the query vector and Y as the key–value vectors, i denotes the index of the layer ranging from 1 to L, $h_{ts}^0 = h_{tc}^0 = h_t$, $\tilde{M}^i$ and $\tilde{T}^i$ are the intermediate feature in the ith transformer block, h_{ts}^{i-1} and h_{tc}^{i-1} denote the in-sample knowledge and cross-sample knowledge in the ith layer, respectively. We can get the estimate of $\hat{M} \approx h_{ts}^L$ and $\hat{T} \approx h_{tc}^L$. We use a multilayer perceptron to fuse in-sample knowledge $\hat{M}$ and cross-sample knowledge $\hat{T}$ to obtain the causality-aware representation of the text modality [16].

3.3.3. Counterfactual cross-modal attention module

Traditional cross-modal attention learns to map relationships between modalities by maximizing the likelihood. The central interaction mediator, namely the cross-modal attention score, remains unexplored, resulting in an unknown mapping relationship between modality representations and fusion outcomes [30]. To evaluate the causal effects of modality and attention on fusion results to achieve causality-aware multimodal fusion **(referred to as Q2 in the Introduction)**, we turn to the causal effect framework [43] as a basis for reimagining the cross-modal attention mechanism [24].

Theoretical Analysis of Cross-Modal Causality Discovery. Causal inference allows us to open the black box of attention and teach models to distinguish between main and side effects. To build a robust fusion mechanism, we first decouple cross-modal attention. Cross-modal attention presently represents the predominant approach for modal fusion, which originates from the following self-attention formula:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{14}$$

where the query vector Q, key vector K, and value vector V come from the same modality. However, in cross-modal attention, one modality is designated as K and V, while the other is designated as Q. This mechanism integrates information by analyzing the mapping relationship between the two modalities.

Motivated by counterfactual theory [18], a straightforward approach is to assess the causal impact of attention scores on fusion outcomes. Concretely, we devise a counterfactual scenario to understand "bad attention". By training the model under the guidance of this counterfactual attention scenario, the model acquires better attention mechanisms in the factual scenario for fusion.

For further analysis, we use uppercase letters (*e.g.* Q) to represent random variables, lowercase letters such as (*e.g.* m) to represent observed values of random variables, and subscripts (*e.g.* F_x) to describe the effect of the action $do(X = x)$ on F for further analysis, as illustrated in Fig. 5. Q, K, and V represent the aforementioned branch vectors used to calculate attention, A denotes the attention scores, and F represents the fusion result as illustrated in Fig. 5(a).

In the factual scenario, we assign K and V to one modality, denoted as m_1, and designate Q as the other modality, represented by m_2, akin to the conventional cross-modal attention setup, yielding the causal graph as shown in Fig. 5(b). In the counterfactual scenario, K and V will be replaced with a value (*e.g.* m_1^*) different from m_1. Consequently, a series of "what if" questions arise. For example, $F_{m_1,A_{m_1^*,m_2}}$ represents "What would the fusion result be if K and V were set to m_1 and A were set to the value when K and V were m_1^* and Q were set to m_2?" in Fig. 5(c); $F_{m_1^*,A_{m_1^*,m_2}}$ represents "What would the fusion result be if K and V were set to m_1^* while Q were set to m_2" in Fig. 5(d).

Causal effect specifically quantifies the change in the outcome variable resulting from the manipulation or change in the causal variable while holding other relevant factors constant [43,44]. Specifically, the total effect (TE) refers to the overall impact of a variable on another variable. It encompasses both direct (*e.g.* $KV \rightarrow F$) and indirect effects (*e.g.* $KV \rightarrow A \rightarrow F$). Suppose we want to measure the total effect of K and V on the fusion result F, we can compare the difference in the fusion result from changing situation $do(KV = m_1^*)$ to situation $do(KV = m_1)$ [43]:

$$E_{TE}^{m_1} = F_{m_1,A_{m_1,m_2}} - F_{m_1^*,A_{m_1^*,m_2}}, \tag{15}$$

where $E_{TE}^{m_1}$ denotes the total effect of m_1 on the fusion result F, $F_{m_1,A_{m_1,m_2}}$ denotes the traditional factual fusion result, $F_{m_1^*,A_{m_1^*,m_2}}$ denotes the fusion result when we set K and V to m_1^*. We further discuss the natural direct effect (NDE), which represents the expected change in F when changing situation $do(KV = m_1^*)$ to situation $do(KV = m_1)$ that is not mediated by any variables. In situation $do(KV = m_1^*)$, the value of A is $A_{m_1^*,m_2}$. Keeping A unchanged, the NDE of K and V on the fusion outcome F is expressed as follows:

$$E_{NDE}^{m_1} = F_{m_1,A_{m_1^*,m_2}} - F_{m_1^*,A_{m_1^*,m_2}}, \tag{16}$$

where $F_{m_1,A_{m_1^*,m_2}}$ denotes the fusion result when we set K and V to m_1 and set A to the attention scores interacting between m_1^* and m_2. Meanwhile, the total indirect effect (TIE) measures the difference between TE and NDE, indicating how K and V impact the fusion outcome via the intermediary of attention scores A:

$$E_{TIE}^{m_1} = E_{TE}^{m_1} - E_{NDE}^{m_1} = F_{m_1,A_{m_1,m_2}} - F_{m_1,A_{m_1^*,m_2}}, \tag{17}$$

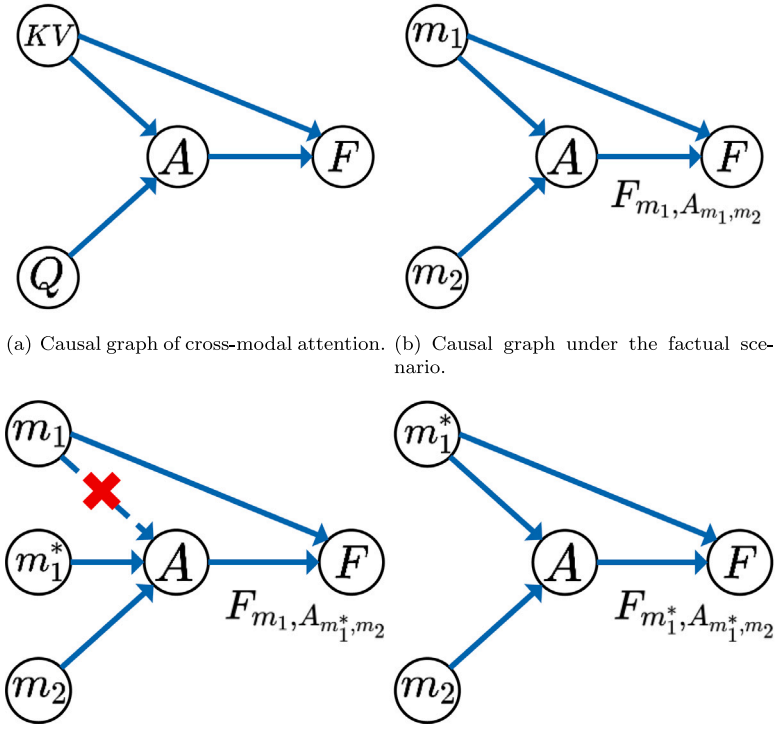
we observe that the TIE of K and V on the fusion outcome essentially corresponds to the TE of attention scores on the fusion outcome, denoted as E_{TE}^A.

Implementation with the deep neural network of CCoAt. In the above process, we do not intervene with attention scores or the fusion result with Q, and Q is exogenous in the causal graph, we can treat it as a constant and omit it from the equation to streamline the network architecture:

$$E_{TE}^A = E_{TIE}^{m_1} = F_{m_1,A_{m_1}} - F_{m_1,A_{m_1^*}}. \tag{18}$$

Based on the conclusion derived from Eqs. (17) and (18), we can employ a neural network to assess the indirect effect effects between the modality m_1 and the outputs. Meanwhile, this network can assess the total effects between the attention scores and the outputs.

By employing counterfactual interventions, we prompt the model to speculate about attention maps that do not exist and explore their



(a) Causal graph of cross-modal attention. (b) Causal graph under the factual scenario.

(c) Causal graph under counterfactual scenario when KV is set to m_1 and A is set to the value it would have been if KV were m_1^*.

(d) Causal graph when intervening KV to be m_1^*.

Fig. 5. Causal graph of cross-modal attention and counterfactual intervention annotations.

impact on fusion outcomes. Inspired by Eq. (18), we use $\bar{A}$ to represent $A_{m_1^*}$ for simplicity, and introduce do-operator $do(\cdot)$ for intervention [45]. Consequently, we obtain:

$$F(do(A=\bar{A}), KV=m_1) = q_\theta([m_1 * \bar{A}_1; \cdots; m_1 * \bar{A}_h]), \tag{19}$$

where q_θ denotes the fusion network, h is the number of the head in multi-head attention. To implement counterfactual intervention using a deep learning framework [30,32], we rewrite Eq. (18) as follows:

$$\begin{aligned} &E_{TE}^{A} = E_{TIE}^{m_1} = \\ &\mathbb{E}_{\bar{A}\sim\tau}[\underbrace{F(A=A_{m_1}, KV=m_1)}_{CoAt} - \underbrace{F(do(A=\bar{A}), KV=m_1)}_{CCoAt}], \end{aligned} \tag{20}$$

where E_{TE}^{A} and $E_{TIE}^{m_1}$ denote the total effect of the attention A and the total indirect effect of m_1 on the outcome, respectively and τ is the distribution of counterfactual attention which determined by the counterfactual type as discussed in 4.4. The $CoAt(\cdot)$ and $CCoAt(\cdot)$ represent cross-modal attention and counterfactual cross-modal attention, respectively. The formula is as follows:

$$F_{tm} = CoAt_{m\to t} = Encoder(Q=F_t, K=F_m, V=F_m), \tag{21}$$

$$F_{tm}^{*} = CCoAt_{m\to t} = CEncoder(Q=F_t, K=F_m, V=F_m), \tag{22}$$

where F_{tm} denotes the fusion result of $m \in \{a, v\}$ and t, $Encoder(\cdot)$ is an L-layer Transformer that includes cross-attention and feed-forward neural networks, where the text F_t serves as the query vector, and the modalities $F_m, m \in \{v, a\}$, serve as the key and value vectors. $CEncoder(\cdot)$ denotes a counterfactual cross-modal attention block illustrated in the lower right subplot of Fig. 4. $CEncoder(\cdot)$ is architecturally identical to the $Encoder$, with the only differences lying in the attention computations across various layers. The attention of the $CEncoder(\cdot)$ is constrained to be equal to $\bar{A}$, depending on the type of counterfactual attention as discussed in 4.4. F_{tm}^{*} denotes the counterfactual fusion result of $m \in \{a, v\}$ and t. The CCoAt module is only used during training, so there is no additional computational consumption for inference.

3.4. Prediction and loss function

We use a multilayer perceptron to fuse F_{ta}, F_{tv} to get the result of prediction $\hat{y}$ and F_{ta}^{*}, F_{tv}^{*} to obtain the result of the counterfactual prediction $\hat{y}_{cf}$, respectively:

$$\hat{y} = MLP([F_{ta}; F_{tv}]), \tag{23}$$

$$\hat{y}_{cf} = MLP([F_{ta}^{*}; F_{tv}^{*}]). \tag{24}$$

The causal effect of attention on the prediction can be estimated using $\hat{y}$ and $\hat{y}_{cf}$:

$$y_{effect} = \hat{y} - \hat{y}_{cf}, \tag{25}$$

where y_{effect} reflects the gap between factual attention and counterfactual attention, with better factual attention indicating a larger gap. So y_{effect} can serve as a supervision signal during the training process:

$$\mathcal{L}_{effect} = \mathcal{L}_{dis}(y_{effect}, y), \tag{26}$$

where $\mathcal{L}_{dis}$ denotes the distance loss function, and y denotes the ground truth label. For multimodal sentiment analysis, $\mathcal{L}_{dis}$ is the Mean Absolute Error (MAE) loss, and for multimodal humor detection, it is the cross-entropy loss. Besides, we use the hyperparameter γ to balance the counterfactual guidance loss and the original task loss $\mathcal{L}_{task}$. Our total loss function is formulated as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{task} + \gamma\mathcal{L}_{effect} \\ &= \mathcal{L}_{dis}(\hat{y}, y) + \gamma\mathcal{L}_{effect}, \end{aligned} \tag{27}$$

Our proposed AtCAF algorithm is shown in Algorithm. 1.

Algorithm 1: **A**ttention-based **C**ausality-**A**ware **F**usion network (AtCAF)

Input : $D = \{(\mathbb{M}_t, \mathbb{M}_a, \mathbb{M}_v), Y\}$, γ, learning rate η_{main}, η_{bert}, N_{epochs}, global dictionary size N
Output: Prediction $\hat{y}$ # *sentiment score*

```
# Build the confounder dictionary (refer to Section 3.3.2)
h_g ← KMeans(Pool(Transformer(M_t)), N)
for each epoch do
    for mini-batch {(I_t^i, I_a^i, I_v^i), Y_i}_{i=1}^B from D do
        # Unimodal feature extraction (refer to Section 3.3.1)
        F_a, F_v, h_t ← Encode(I_a^i, I_v^i, I_t^i)
        if training then
            # Text debiasing using CATDM (refer to Section 3.3.2)
            F_t ← CATDM(h_t, h_g)
            # Multimodal fusion with CCoAt (refer to Section 3.3.3)
            F_ta ← CoAt_{a→t}(F_a, F_t) # text-audio fusion
            F_tv ← CoAt_{v→t}(F_v, F_t) # text-vision fusion
            F*_ta ← CCoAt_{a→t}(F_a, F_t) # text-audio counterfactual fusion
            F*_tv ← CCoAt_{v→t}(F_v, F_t) # text-vision counterfactual fusion
            # Prediction (refer to Section 3.4)
            ŷ ← MLP([F_ta; F_tv]) # fusion result
            ŷ_cf ← MLP([F*_ta; F*_tv]) # counterfactual fusion result
            # Loss function and optimization (refer to Section 3.4)
            L_effect ← L_dis(y_effect, y) # effect loss
            L ← L_dis(ŷ, y) + γ L_effect # total loss
            Mini-batch gradient descent
            Update model parameters
        else
            F_ta ← CoAt_{a→t}(F_a, F_t) # text-audio fusion
            F_tv ← CoAt_{v→t}(F_v, F_t) # text-vision fusion
            ŷ ← MLP([F_ta; F_tv]) # fusion result
```

3.5. CATDM and CCoAt: Enhancing multimodal sentiment analysis

Traditional approaches often struggle to capture nuanced sentiment cues due to a bias in multimodal contexts. The Context-Aware Textual Data Model (CATDM) tackles this challenge through in-sample and cross-sample knowledge extraction, reducing spurious correlations. Integrating these knowledge extraction methods encourages the model to evaluate knowledge stability, incorporating global dictionary knowledge. For instance, the term "movie" appears in various contexts like actor descriptions, plot summaries, and intonation. Through supervised learning, the model recognizes that "movie" is contextually sensitive, lacking causality-awareness. In contrast, phrases like "nice movie" consistently evoke positive sentiment, establishing "nice" as a stable sentiment indicator. Leveraging these insights, CATDM enhances textual data's reliability for multimodal fusion.

Effective attention aggregate is crucial for interpreting sentimental clues in diverse modalities. Traditional methods often struggle with low-quality attention, requiring human intervention for improvement [30]. CCoAt addresses this issue by posing a critical question: "What would happen to fusion results if the attention was altered to $\bar{A}$". CoAt extracts invariant information, such as causality-aware clues explaining primary causation. Additionally, CCoAt captures unstable contextual information, particularly in volatile multimodal contexts with side effects. Supervised by the loss function, this approach ensures that CoAt's fusion outcomes differ significantly from those with low-quality attention. This optimization enhances cross-attention in CoAt and encourages autonomous prioritization of primary causal effects, reducing the need for human oversight.

Table 2
Dataset splits for CMU-MOSI, CMU-MOSEI, UR-FUNNY, CMU-MOSEI*, and CMU-MOSEI**

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16,326	1871	4659	22,856
CMU-MOSEI*	16,957	1848	1715	20,520
CMU-MOSEI**	16,770	1833	1955	20,558
UR-FUNNY	7614	980	994	9588

4. Experiments

This section presents the experimental details, encompassing the dataset, experimental setup, baseline models, and ablation studies.

4.1. Datasets

We conduct experiments on four public datasets, CMU-MOSI [19], CMU-MOSEI [20] and UR-FUNNY [21], to validate the effectiveness of our proposed AtCAF. To further verify the model's robustness on Out-Of-Distribution (OOD) data, we conduct experiments using the OOD version of the CMU-MOSEI dataset [33]. The data splitting followed the original dataset specifications, detailed in Table 2.

CMU-MOSI. The CMU-MOSI dataset is designed for multimodal sentiment analysis. It encompasses 2199 individual speech segments extracted from 93 opinion-based videos featuring 89 YouTube movie reviewers. It incorporates speech, text transcriptions, and visual clues, offering diverse modalities for sentiment assessment. Annotations include sentiment labels and continuous dimensions ranging from −3 (strongly negative) to 3 (strongly positive).

CMU-MOSEI. The CMU-MOSEI dataset is an expanded version of CMU-MOSI. It captures diverse expressions of emotion, speech, and language within speeches, encompassing a spectrum of natural sentimental experiences. It comprises 3228 videos covering 250 varied topics, sourced from 1000 different YouTube speakers. Data annotation standards are consistent with CMU-MOSI.

CMU-MOSEI-OOD. The CMU-MOSEI-OOD is constructed through an adapted simulated annealing algorithm [46] from the CMU-MOSEI dataset, which iteratively adjusts testing distributions to achieve significant differences in word-sentiment correlations from the training set. It is an OOD dataset in multimodal sentiment analysis. The dataset encompasses two configurations: the OOD distribution for the binary classification scenario (marked as CMU-MOSEI*) and the OOD distribution for the seven-class classification scenario (marked as CMU-MOSEI**), corresponding to these two distinct granularities of data partitioning. The data annotation is consistent with CMU-MOSEI.

UR-FUNNY. The UR-FUNNY dataset comprises 8257 instances of humor extracted from TED talk videos and transcripts. Each instance includes textual, visual, and acoustic modalities for multimodal humor detection (MHD). Binary labels within the UR-FUNNY dataset indicate whether a joke is humorous. The UR-FUNNY task encompasses the context setting (given the context) and the punchline setting (without the context). Our experimentation on UR-FUNNY is conducted within the punchline setting.

4.2. Evaluation metrics

Based on prior research [12], we comprehensively analyze the CMU-MOSI and CMU-MOSEI datasets, reporting metrics pertinent to regression and classification tasks.

Regression Task Metrics

- **Mean Absolute Error (MAE):** This measures the average magnitude of the errors between predicted and actual sentiment scores. The value range is from 0 to infinity, with a lower value indicating a better model fit. MAE is chosen for its simplicity and direct interpretation, representing the average prediction error in the same units as the sentiment scores.

- **Pearson Correlation Coefficient (Corr):** This metric quantifies the strength and direction of the linear relationship between predicted and actual sentiment scores. The value range is from −1 to 1, where 1 indicates a perfect positive linear relationship, −1 indicates a complete negative correlation, and 0 indicates no linear relationship. Corr is selected for its ability to represent the degree of linear association, which is essential for understanding the model's predictive power.

Classification Task Metrics

- **Seven-category Classification Accuracy (ACC-7):** ACC-7 divides the range from −3 to 3 into seven score intervals to assess the model's accuracy in fine-grained sentiment analysis. The value range is from 0 to 1. High values indicate a better model performance in fine-grained sentiment classification.
- **Binary Classification Accuracy (ACC-2):** This measures the proportion of correct predictions in a binary classification context. It has the same value range and implications as ACC-7.
- **Weighted F1-score (F1):** This balances precision and recall for classification tasks, with different weights for classes to account for imbalance. The value range is from 0 to 1, with a higher value indicating a better balance between precision and recall. F1 is chosen for its ability to provide a single measure that accounts for false positives and false negatives, which is crucial for models that balance these two types of errors. Specifically, for ACC-2 and F1, we report the accuracy under the following two configurations.

    1. **Non-negative/Negative (NN/N) Classification:** This evaluates the model's ability to distinguish between non-negative (≥ 0) and negative (<0) sentiments.
    2. **Positive/Negative (P/N) Classification:** This focuses on the model's accuracy in classifying sentiments as either positive (>0) or negative (<0), which is essential for understanding the model's performance on the primary sentiment classes.

Specifically, For CMU-MOSEI*, we report ACC-2 and F1, while for CMU-MOSEI**, we report ACC-7 with their settings consistent with those in CMU-MOSEI. It is worth mentioning that the ACC-2 and F1-score on the OOD dataset not only reflect the model's predictive accuracy but also indicate the model's robustness. In addition, we report ACC-2 as the evaluation metric to assess the model's accuracy in identifying humor on the UR-FUNNY dataset. The value range is from 0 to 1. High values signify greater accuracy of the model in identifying humor and vice versa.

4.3. Baselines

We compare our model with other MSA models to validate the performance of our causality-aware sentiment analysis network. To ensure a fair comparison, all baseline models employ the same task-specific loss function and utilize the same experiment settings. The baselines are listed as follows:

TFN [3] utilizes tensors at three levels—unimodal, bimodal, and trimodal for modality fusion.

LMF [4] generates a multimodal output representation through low-rank multimodal fusion involving modality-specific factors.

MulT [6] employs cross-modal attention between each pair of the three modalities for fusion.

MISA [23] projects modalities to invariant and specific subspaces, enabling input reconstruction and fusion for task predictions.

MAG [25] employs multimodal adaptation gates across various transformer layers to adaptively fuse non-textual information.

Self-MM [11] utilizes a modal label generation task to preserve modality-specific information.

MMIM [12] utilizes multi-level mutual information maximization to model the process of modality fusion.

BBFN [47] designs a text-centric Modality Complementation Layer to achieve different modality fusion and separation representations.

HyCon [48] utilizes three different hierarchical levels of contrastive learning to acquire representations of modalities.

CubeMLP [49] employs separate MLP units to independently mix multimodal features at sequential, channel, and modality levels, enabling the blending of these features across all axes.

TETFN [7] empowers non-textual modalities using textual information and constructs a Transformer network to model modality fusion.

AOBERT [13] designs multimodal masking and alignment prediction tasks to perform multimodal fusion in bert-large.

MTMD [50] learns modality interactions by distilling different tasks and modalities, leveraging weighted concatenation.

CRNet [51] designs a multi-task strategy to optimize different representation subspaces.

TMBL [52] proposes a novel modality-binding network to extract modality-invariant and modality-specific information efficiently.

PEST [53] introduces a dynamic propagation model to enhance sentiment analysis through cross-modal interaction.

MCL-MCF [54] employs multi-level contrastive learning and multi-layer convolution fusion for multimodal sentiment analysis.

CLUE [33] evaluates causal effects within models by introducing additional text models, which is a method tailored for OOD scenarios.

4.4. Implement details

Feature Extraction. To ensure a fair comparison, we adopt the data preprocessing approach of the HyCon [48]. Specifically, Facet [37] is applied for visual modality feature extraction, capturing features such as facial action units, landmarks, and head pose, sampled at 30 Hz to sequence facial expressions over time. For the audio modality, COVAREP [36] extracts features including 12 Mel-frequency cepstral coefficients, pitch, speech polarity, and spectral properties, sampled at 100 Hz to track vocal tone changes throughout each utterance. We utilize XLNet [55] for textual embeddings, with an embedding dimension of 768.

Hyperparameter Setting. We utilize Adam as the optimizer with a learning rate of 5e−5 for the text encoder, 1e−3 as the primary learning rate, and maintain a fixed batch size of 64. The lengths of all modalities are padded to the length of the longest sample within the same batch. The hidden dimension of the LSTM is fixed at 32, and the kernel size of the Conv1D is fixed at 1. The hidden dimension for all cross-modal and counterfactual cross-modal attention is 30, with a dropout rate of 0.1 applied to the attention computation and the fully connected layer. The number of heads in the multi-head attention is 5. The tuning ranges for other hyperparameters are delineated as follows: γ is tuned within the set {0.1, 0.2, 0.4, 0.8}, the sizes of the confounder dictionary for text are explored within {50, 100, 200, 400}, the number of transformer blocks in CATDM is varied among {1, 3, 6, 9}, and the number of transformer blocks in CCoAt is examined within {1, 2, 4, 6}. All experiments are executed on a single Tesla V100 (32 GB) GPU equipped with CUDA 11.2, using PyTorch version 1.8.1.

Counterfactual Attention Strategy. We implement four counterfactual attention mechanisms—uniform attention, reversed attention, shuffle attention, and random attention in the CCoAt module [30]. Their implementation details are outlined as follows:

- **Uniform attention:** Unmasked positions are assigned values equivalent to the average of the actual attention weights, ensuring each unmasked token receives the same weight.
- **Reversed attention:** Unmasked positions are assigned values obtained by subtracting the actual attention from the maximum attention score. Consequently, the score assigned to each unmasked token is forcibly reversed compared to the actual attention.
- **Shuffle attention:** Attention scores are shuffled along the batch dimension.
- **Random attention:** Unmasked positions are assigned attention scores uniformly sampled from the distribution $\mathcal{U}(0, 2)$.

Table 3
Comparison results between AtCAF and other baselines on the CMU-MOSI, CMU-MOSEI, and UR-FUNNY. For ACC-2 and F1, the values to the left of "/" represent the results under the non-negative/negative (NN/N) setting. In contrast, the values to the right represent the results under the positive/negative (P/N) setting. Metrics labeled with ↑ indicate better performance with higher values, and those labeled with ↓ indicate better performance with lower values. The previous state-of-the-art (SOTA) values are annotated with underscores, while the current SOTA values are highlighted in bold.

Models	CMU-MOSI					CMU-MOSEI					UR-FUNNY
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	ACC-2↑
TFN(2017)[a] [3]	0.901	0.698	34.9	–/80.8	–/80.7	0.593	0.700	50.2	–/82.5	–/82.1	67.53
LMF(2018)[a] [4]	0.917	0.695	33.2	–/82.5	–/82.4	0.623	0.677	48.0	–/82.0	–/82.1	68.57
MulT(2019)[a] [6]	0.861	0.711	–	81.5/84.1	80.6/83.9	0.580	0.703	–	–/82.5	–/82.3	70.55
MISA(2020)[a] [23]	0.804	0.764	42.3	80.79/82.10	80.77/82.03	0.568	0.724	–	82.59/84.23	82.67/83.97	70.61
MAG-BERT(2020)[b] [25]	0.790	0.768	42.9	–/83.5	–/83.5	0.602	0.778	51.9	–/85.0	–/85.0	–
MAG-XLNet(2020)[b] [25]	0.746	0.804	42.3	–/84.8	–/84.8	0.581	**0.797**	52.7	–/85.4	–/85.4	–
Self-MM(2021)[a] [11]	0.712	0.795	45.79	82.54/84.77	82.68/84.91	0.529	0.767	53.46	82.68/84.96	82.95/84.93	–
MMIM(2021)[a] [12]	0.700	0.800	46.65	84.14/86.06	84.00/85.98	0.526	0.772	54.24	82.24/85.97	82.66/85.94	–
BBFN(2021)[c] [47]	0.776	0.755	45.0	–/84.3	–/84.3	0.529	0.767	54.8	–/86.2	–/86.1	71.68
HyCon-BERT(2022)[b] [48]	0.713	0.790	46.60	–/85.2	–/85.1	0.601	0.776	52.80	–/85.4	–/85.6	–
HyCon-XLNet(2022)[b] [48]	0.688	0.818	46.0	–/85.5	–/85.4	0.590	0.788	53.2	–/86.4	–/86.4	–
CubeMLP(2022)[c] [49]	0.770	0.767	45.5	–/85.6	–/85.5	0.529	0.760	54.9	–/85.1	–/84.5	–
TETFN(2023)[c] [7]	0.717	0.800	–	84.05/86.10	83.83/86.07	0.551	0.748	–	84.25/85.18	84.18/85.27	–
AOBERT(2023)[c] [13]	0.856	0.700	40.2	85.2/85.6	85.4/86.4	0.515	0.763	54.5	84.9/86.2	85.0/85.9	70.82
MTMD(2023)[c] [50]	0.705	0.799	**47.5**	84.0/86.0	83.9/86.0	0.531	0.767	53.7	84.8/86.1	84.9/85.9	–
CRNet(2024)[c] [51]	0.712	0.797	47.4	–/86.4	–/86.4	0.541	0.771	53.8	–/86.2	–/86.1	–
TMBL(2024)[c] [52]	0.867	0.762	36.3	81.78/83.84	82.41/84.29	0.545	0.766	52.4	84.23/85.84	84.87/85.92	–
PEST(2024)[c] [53]	0.723	0.796	–	–/86.1	–/86.1	0.542	0.761	–	–/85.3	–/85.1	–
MCL-MCF(2024)[c] [54]	0.692	0.799	–	84.9/87.3	84.7/87.2	0.536	0.767	–	84.2/86.4	84.4/86.3	–
AtCAF	**0.650**	**0.831**	46.50	**87.03/88.57**	**86.96/88.53**	**0.508**	0.785	**55.85**	**86.03/86.98**	**86.04/86.78**	**72.13**

The result of UR-FUNNY comes from [13] and [47].

[a] Results are from [12].

[b] Results are from [48].

[c] Results are from original papers.

4.5. Performance results

Under the normal setting of the dataset, the comparison results between our model and other baselines on CMU-MOSI, CMU-MOSEI, and UR-FUNNY are presented in Table 3. Our model achieves state-of-the-art (SOTA) performance on almost all metrics on three datasets, surpassing models with even more powerful text encoders.

Traditional tensor-based fusion networks (TFN, LMF, CubeMLP) struggle to integrate multimodal data, resulting in poor performance effectively. The attention-based networks (MulT, MAG-BERT, MAG-XLNet, BBFN, TETFN, CRNet, TMBL, PEST) learn mappings between modalities by considering the similarities and co-occurrences of different modalities, which compensates for the shortcomings of tensor networks and achieves an improvement in accuracy. However, they do not consider the sentimental cues in the fusion process. By combining auxiliary tasks such as modal reconstruction (MISA), unimodal prediction (Self-MM), maximizing mutual information (MMIM), contrastive learning (HyCon-BERT, HyCon-XLNet, MTMD, MCL-MCF), mask prediction (AOBERT), the network model based on multi-task learning has also achieved good performance. However, they still model based on correlation, and the distributional bias inherent in the dataset remains unresolved. Unlike previous works, AtCAF employs a CATDM module to overcome the inherent distribution bias in the dataset, making it easier to mine genuine sentimental cues. During the fusion stage, AtCAF uses CCoAt to filter out incorrect contexts, achieving consistent sentimental cues across different modalities. Thus, AtCAF achieves exceptional performance on the CMU-MOSI dataset across various metrics, notably improving upon previous SOTA values. Specifically, compared to prior benchmarks, the MAE drops from 0.688 to 0.650, the Corr increases from 0.818 to 0.831, the ACC-2 increases from 85.2%/87.3% to 87.03%/88.57%, and the F1 increases from 85.4%/87.2% to 86.96%/88.53%. Experiments have proven that AtCAF can achieve sota performance in CMU-MOSI even better than models trained with larger parameters and more complex training processes. It is worth noting that AtCAF shows significant improvements in MAE, Corr, ACC-2, and F1, which means that AtCAF has a better fit than the baselines and, with the assistance of the debiasing module, possesses excellent sentiment polarity discrimination capabilities.

On the CMU-MOSEI dataset, traditional tensor fusion networks are limited by their scalability and size, making it difficult to achieve good performance. Transformer-based networks have expanded in scale and improved how they model multimodal data, leading to increased accuracy. However, although many models have used an enhanced module when facing more complex contexts, their fusion process is inefficient, and their performance has hit a bottleneck. Even with the use of other auxiliary tasks to reduce the burden of modal fusion, it is impossible to break through this bottleneck, as the correlation-based modeling approach causes it. Unlike other baseline models, AtCAF designs the CATDM from a causal perspective to mitigate situations where spurious correlations easily influence models in correlation-based modeling. Furthermore, during the fusion stage, CCoAt is used to encourage the model to consider not only the co-occurrence of modalities when learning attention scores but also the inherent impact of modeling unstable contexts on attention, enhancing the effect of modal fusion in complex contexts. Thus, AtCAF exhibits superior performance compared to all baselines. The richer multimodal contexts provided by CMU-MOSEI prove advantageous for AtCAF's CATDM and CCoAt modules in uncovering genuine causal relationships. In comparison to prior SOTA values, AtCAF achieves new SOTA benchmarks across most metrics, with a notable decrease in MAE by 0.07, an improvement in ACC-7 by 0.95%, and an increase in ACC-2 and F1 by 1.03%/0.58% and 1.04%/0.38%, respectively. These experiments affirm AtCAF's capability to discern sentimental cues accurately within a multimodal context. It is worth mentioning that AtCAF has shown significant improvement on ACC-7, ACC-2, and F1, which implies that AtCAF is particularly advantageous for multi-modal sentiment analysis of different granularities on larger datasets.

Furthermore, on the UR-FUNNY dataset, tensor-based networks and attention-based networks both perform poorly because multimodal humor detection is a more abstract and challenging problem. AOBERT uses the more powerful bert-large to mine text information, and BBFN compensates for the supplementary information during multimodal interaction through multimodal complement layers. Despite this, AtCAF

Table 4
Comparison results of AtCAF and other baseline models on the out-of-distribution datasets CMU-MOSEI* and CMU-MOSEI**. The labeling and settings are consistent with those in Table 3.

Models	CMU-MOSEI*		CMU-MOSEI**
	ACC-2↑	F1↑	ACC-7↑
TFN [3]	71.23/69.76	70.46/69.02	41.05
LMF [4]	68.16/69.58	68.31/69.58	31.11
MulT [6]	72.56/73.73	72.44/73.58	40.58
MAG-BERT [25]	74.59/76.41	74.48/76.27	45.88
MISA [23]	74.48/76.45	74.39/76.33	43.15
Self-MM [11]	74.68/74.50	74.33/74.22	45.81
MAG-BERT+CLUE [33]	78.34/80.51	78.23/80.46	48.66
MISA+CLUE [33]	77.17/78.77	77.08/78.74	46.86
SELF-MM+CLUE [33]	77.76/79.48	77.72/79.47	48.09
AtCAF	**78.60/80.62**	**78.59/80.60**	**50.13**

All results are from [33].

still performs best with the causality-aware representations provided by CATDM and the context-aware capabilities offered by CCoAt. AtCAF elevates binary classification accuracy by 0.45%, underscoring the contribution of causal information in multimodal humor detection.

Under the out-of-distribution setting, the comparative results between AtCAF and the baselines on CMU-MOSEI are shown in Table 4. When operating under the out-of-distribution (OOD) setting, excessive reliance on information from the textual modality or the absence of an effective method for modal fusion can significantly diminish the model's accuracy in both binary and seven-category classification. Traditional methods (TFN, LMF, MulT, MAG-BERT, MISA, Self-MM) suffer a significant reduction in accuracy. CLUE introduces an additional text model from the perspective of measuring causal effects to alleviate the aforementioned issues and enhance the ability of fundamental multimodal sentiment analysis models to overcome out-of-distribution issues. However, it cannot mine causality across modalities. Therefore, it demonstrates the potential of causal inference in uncovering sentimental cues but does not fully utilize it. Building on the causal representation of the text modality using CATDM, the CCoAt is applied to explore causal cues between different modalities. This allows AtCAF to improve the current SOTA values under the out-of-distribution setting, with an increase of 0.26%/0.11% in ACC-2, 0.36%/0.14% in F1, and 1.47% in ACC-7. Experiments under the out-of-distribution setting confirm that AtCAF can capture more stable sentimental cues than other models. Even when there is a significant label shift between the training and testing sets, optimal results are still achieved in classification problems of different granularities. This also verifies the robustness of CATDM and CCoAt in capturing sentimental cues. It is worth noting that AtCAF has made significant improvements on ACC-7, indicating that AtCAF is particularly advantageous for fine-grained multi-modal sentiment analysis on the OOD datasets.

In summary, the AtCAF achieves new state-of-the-art values across all datasets in both normal and out-of-distribution (OOD) settings, including CMU-MOSI, CMU-MOSEI, UR-FUNNY, CMU-MOSEI*, and CMU-MOSEI** without notable drawbacks. Its performance remains superior to all models, even without employing complex training processes such as contrastive learning or utilizing larger text encoders. It confirms the effectiveness and robustness of the CATDM and CCoAt modules in multimodal sentiment analysis.

4.6. Ablation studies

In this section, we conduct ablation experiments to validate the effectiveness of our proposed module, including the CATDM and CCoAt modules. The experimental results on the CMU-MOSI, CMU-MOSEI, and UR-FUNNY under normal settings are shown in Table 5. In contrast, the results on the CMU-MOSEI under the out-of-distribution setting are shown in Table 6.

The effectiveness of CATDM. To validate the effectiveness of CATDM, we conduct experiments where CATDM is removed (denoted as w/o CATDM) and where k-means initialization for CATDM is omitted (denoted as w/o CATDM_{init}).

Under normal settings, we observe a decrease in accuracy across all datasets after removing CATDM. This decline in performance underscores the pivotal role of debiasing for textual modalities in facilitating modal fusion and enhancing the overall network's performance. Furthermore, we investigate the impact of different initialization strategies for CATDM. Surprisingly, we find that model performance improves even when CATDM parameters are randomly initialized, suggesting the presence of confounders in the data. However, initializing the global dictionary using causal inference methods yields superior results.

We can arrive at consistent conclusions on the out-of-distribution dataset, which is particularly evident in the binary classification task of negative/positive discrimination. Removing the k-means initialization directly results in an approximate 3% decrease, while directly removing CATDM leads to an approximate 4% decrease. Due to the OOD settings, distinguishing between negative and positive sentiments becomes more vulnerable to the influence of confounders.

The observations above confirm that CATDM can effectively remove biases from both normal and out-of-distribution data, obtaining causal representations of the text. The front-door adjustment effectively mitigates the spurious correlations introduced by confounders, thereby enhancing the robustness and efficacy of the model.

The effectiveness of CCoAt. To verify the role of the counterfactual cross-modal attention module, we remove the counterfactual attention module for the fusion of text and vision (denoted as CCoAt_{ta}), the counterfactual attention module for the fusion of text and audio (denoted as CCoAt_{tv}), and remove all counterfactual attention modules (denoted as CCoAt), respectively. To further explore the role of counterfactual strategies in cross-modal counterfactual attention, we have built three suboptimal variants based on retaining the corresponding modules. For instance, CCoAt^{S}_{ta} denotes the variant that retains the structure of the text-audio branch, but rather than sampling attention scores from a counterfactual distribution, they are directly optimized from the loss function.

Under normal settings, when all counterfactual modules are removed, the accuracy decreases the most. Specifically, ACC-2 decreases by 3.21% on CMU-MOSI, 1.59% on CMU-MOSEI, and 3.02% on UR-FUNNY. This suggests that counterfactual modules can enhance cross-modal attention learning, improving modal fusion and model accuracy. Furthermore, on the CMU-MOSI and UR-FUNNY datasets, removing the text-audio CCoAt module leads to a more pronounced decrease in model accuracy than removing the text-vision CCoAt module. Conversely, the opposite trend is observed in the CMU-MOSEI dataset. This disparity suggests that different datasets harbor distinct sentimental cues, influencing the modality fusion of CoAt and the contextual filtering of CCoAt. By comparing the performance of CCoAt_{ta} and CCoAt^{S}_{ta}, CCoAt_{tv} and CCoAt^{S}_{tv}, CCoAt and CCoAt^{S} across the three datasets, it can be observed that even without sampling, i.e., by optimizing directly through the loss function, the corresponding branches can also enhance the model's performance. The reason is that, without sampling from the counterfactual distribution, the corresponding branches will explore contexts that hinder the stable prediction of CoAt according to the loss function, thus serving the same purpose of uncovering unstable contexts. However, comparing the complete model reveals that these suboptimal models do not perform as well as the full AtCAF. This is because the counterfactual strategy endows CCoAt with a prior distribution, which enhances CCoAt's ability to mine unstable contexts. Moreover, the more branches in CCoAt that employ counterfactual strategies, the better the model's performance improvement, confirming the effectiveness of counterfactual strategies. The performance of the two variants when a single-branch counterfactual strategy is removed is also influenced by the properties of the data.

Table 5
The ablation studies on the CMU-MOSI, CMU-MOSEI, and UR-FUNNY datasets. The terms 'w/o CATDM', 'w/o CCoAt', and 'w/o ALL' represent the removal of the CATDM module, all CCoAt modules, including the text-audio branch and the text-vision branch, and both the CATDM and CCoAt modules, respectively. 'w/o CATDM_{init}' indicates not using k-means to initialize the global dictionary. The terms 'w/o CCoAt_{ta}' and 'w/o CCoAt_{tv}' denote the removal of the text-audio counterfactual learning branch or the text-vision counterfactual learning branch within CCoAt. 'w/o CCoAt_{ta}^{S}', 'w/o CCoAt_{tv}^{S}', 'w/o CCoAt^{S}' indicate that we have retained the corresponding part of the model architecture, but have not employed any counterfactual strategy. Instead, it learns directly from the loss function. It can be considered a suboptimal variant compared to the complete component removal.

Models	CMU-MOSI					CMU-MOSEI					UR-FUNNY
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	ACC-2↑
AtCAF	**0.650**	**0.831**	46.50	**87.03/88.57**	**86.96/88.53**	**0.508**	**0.785**	**55.85**	86.03/**86.98**	86.04/**86.78**	**72.13**
w/o CATDM	0.681	0.820	46.65	84.55/87.04	84.41/86.98	0.528	0.774	53.96	85.30/86.41	85.39/86.25	69.52
w/o CATDM_{init}	0.669	0.821	46.50	85.42/87.04	85.40/87.06	0.525	0.773	53.10	**86.11**/86.74	**86.11**/86.53	70.22
w/o CCoAt_{ta}	0.694	0.815	46.50	84.11/85.98	84.04/85.96	0.520	0.784	54.60	85.02/86.57	85.10/86.38	69.72
w/o CCoAt_{ta}^{S}	0.675	0.819	46.21	85.13/87.50	85.02/87.46	0.517	0.777	54.95	85.23/86.54	85.39/86.44	71.03
w/o CCoAt_{tv}	0.700	0.820	45.63	84.40/86.74	84.20/86.62	0.518	0.774	54.88	84.95/86.46	85.07/86.30	71.23
w/o CCoAt_{tv}^{S}	0.697	0.819	46.65	84.84/86.42	84.81/86.44	0.520	0.784	53.72	85.40/86.49	85.47/86.31	71.33
w/o CCoAt	0.661	0.824	46.50	83.82/85.67	83.78/85.68	0.529	0.765	54.99	84.44/86.79	84.70/86.71	69.11
w/o CCoAt^{S}	0.686	0.807	**47.08**	84.69/86.43	84.66/86.44	0.518	0.772	54.78	85.90/85.80	85.67/85.40	70.52
w/o ALL	0.697	0.813	44.90	83.53/86.13	83.36/86.05	0.523	0.777	54.07	83.77/86.27	84.06/86.19	69.11

Table 6
The ablation studies on the out-of-distribution datasets CMU-MOSEI* and CMU-MOSEI**. The model variant labels are consistent with those in Table 5.

Models	CMU-MOSEI*		CMU-MOSEI**
	ACC-2↑	F1↑	ACC-7↑
AtCAF	**78.60/80.62**	**78.59/80.60**	**50.13**
w/o CATDM	77.14/76.02	76.78/75.80	48.18
w/o CATDM_{init}	77.67/77.73	77.43/77.59	48.59
w/o CCoAt_{ta}	76.85/76.33	76.65/76.22	48.69
w/o CCoAt_{ta}^{S}	77.08/78.36	77.17/78.39	48.90
w/o CCoAt_{tv}	77.61/76.88	77.26/76.66	47.83
w/o CCoAt_{tv}^{S}	78.37/76.88	77.93/76.62	49.72
w/o CCoAt	75.86/75.08	75.56/74.88	47.16
w/o CCoAt^{S}	77.14/76.80	76.89/76.64	48.66
w/o ALL	74.40/75.31	74.41/75.29	47.01

Table 7
Sensitive analysis of γ.

γ	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
0.1	0.718	85.28/86.89	**0.508**	86.03/**86.98**
0.2	0.669	84.99/86.13	0.526	**86.13**/86.82
0.4	**0.650**	**87.03/88.57**	0.524	85.49/85.61
0.8	0.718	84.40/87.04	0.518	85.81/86.38

Table 8
Sensitive analysis of the number of transformer blocks in CATDM.

L_{CATDM}	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
1	0.707	85.57/87.80	0.544	85.81/85.17
3	**0.650**	**87.03/88.57**	**0.508**	86.03/**86.98**
6	0.696	85.28/87.50	0.527	84.35/86.52
9	0.677	85.28/86.74	0.531	**86.07**/86.35

On the out-of-distribution dataset, the conclusions are consistent with those in normal settings: when CCoAt is completely removed, the model's performance in both binary and seven-category classification is the worst. Even without counterfactual strategies, specific branches can learn to filter contexts through the loss function. However, their performance is still not as good as that of models with counterfactual prior distributions.

In summary, the analysis underscores the importance of the CCoAt module in efficiently filtering multimodal contexts and guiding CoAt to aggregate sentimental information effectively for fusion under normal and OOD settings. Furthermore, employing counterfactual strategies in CCoAt enhances its ability to capture unstable contexts, thereby improving the model's performance. Additionally, compared to the normal setting, removing specific components on OOD data leads to a more pronounced decrease in accuracy, especially in discriminating positive/negative sentiments. This indicates that CATDM and CCoAt are particularly important for sentiment analysis on OOD data.

4.7. Sensitivity and quantitative analysis

In this section, we conduct sensitivity analyses of hyperparameters and a quantitative analysis of the properties of different types of counterfactual attention.

We analyze the impact of the hyperparameter γ on the model's performance. γ reflects the guidance strength of the CCoAt module on the CoAt module. The results are presented in Table 7. We find that the optimal γ is 0.4 on the CMU-MOSI dataset, while on the CMU-MOSEI dataset, it is 0.1. This indicates that the γ plays a role in balancing the guidance strength and label supervision, and the guiding strength of CCoAt needs to be appropriately adjusted based on the differences in datasets. Moreover, these findings underscore the indispensability of the CCoAt module in modulating the guidance strength effectively, thereby enhancing model performance across diverse datasets.

We conduct a sensitivity analysis on the number of transformer blocks in CATDM, as depicted in Table 8. Notably, we observe a close correlation between our model's performance and the number of layers in CATDM. Interestingly, our model achieves optimal performance on both datasets when the number of layers is set to 3. Upon further examination, we deduce that the number of layers in CATDM influences the incorporation of global information into the modality fusion process. As the number of layers increases, more extensive global information is introduced. While this can enhance the model's ability to mitigate spurious correlations by appropriately incorporating cross-sample knowledge, excessive intra-sample and cross-sample information integration during modality fusion can lead to decreased accuracy. This finding underscores the delicate balance required in modality fusion. Appropriately incorporating cross-sample knowledge can facilitate the acquisition of better unimodal representations for fusion. However, excessive integration of such knowledge may burden the modality fusion process, ultimately diminishing accuracy.

Additionally, we conduct a sensitivity analysis on the number of layers in the CCoAt module, as depicted in Table 9. Our findings reveal that employing two layers of Transformer blocks to model counterfactual attention yields optimal results on both datasets. Moreover, we observe a significant decrease in model efficacy when the number of layers in CCoAt exceeds an optimal threshold. This decline in performance can be attributed to the instability introduced in the output of CCoAt (y_{cf}) due to incremental layers of counterfactual intervention. Consequently, CoAt encounters challenges in effectively filtering modal fusion clues based on the information provided by CCoAt, resulting in

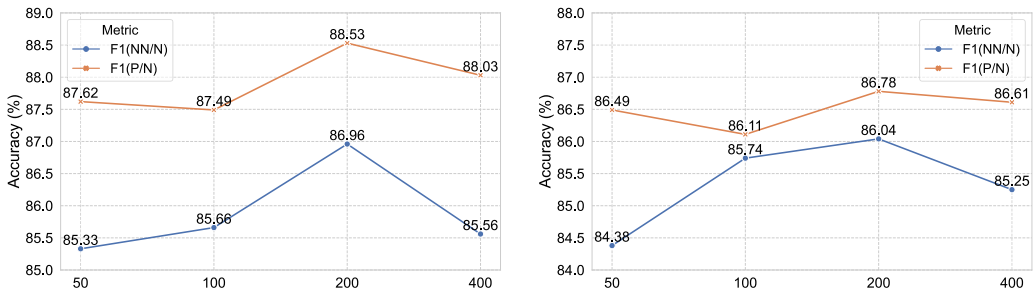


Fig. 6. The sensitivity analysis of the global dictionary size N on the F1-score of the AtCAF.

Table 9
Sensitive analysis of the number of transformer blocks in CCoAt.

L_{CCoAt}	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
1	0.712	84.26/86.59	0.521	85.04/86.68
2	**0.650**	**87.03/88.57**	**0.508**	**86.03/86.98**
4	0.684	84.69/87.20	0.523	83.97/85.86
8	0.694	84.55/85.98	0.531	85.53/85.86

Table 10
Quantitative analysis of different types of attention in CCoAt.

Type	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
Uniform	**0.650**	**87.03/88.57**	0.530	85.34/85.69
Reversed	0.689	84.26/86.43	**0.508**	**86.03/86.98**
Shuffle	0.676	85.13/86.74	0.527	85.73/85.88
Random	0.718	85.28/86.89	0.519	84.85/86.27

diminished accuracy. In summary, the optimal deployment of counterfactual cross-modal attention aids in enhancing the processing of causal clues within cross-modal attention mechanisms.

We perform a quantitative analysis of various counterfactual attention mechanisms, as presented in Table 10. Our observations reveal that different counterfactual attention mechanisms exhibit distinct performances across datasets. Specifically, uniform attention demonstrates superior results on the CMU-MOSI dataset, whereas reversed attention outperforms others on the CMU-MOSEI dataset. This discrepancy underscores the notion that different datasets may necessitate utilizing specific types of counterfactual attention to guide CoAt effectively. These findings reaffirm the role of counterfactual attention in facilitating CoAt's acquisition of causality-aware sentimental clues.

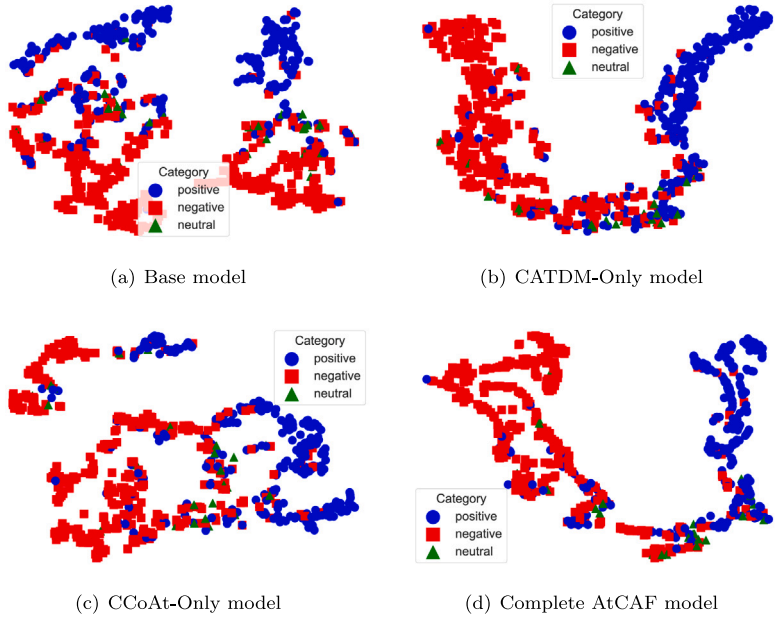
Finally, we analyze the relationship between the global dictionary size N and the performance of AtCAF as illustrated in Fig. 6. The x-axis represents the dictionary size, and the y-axis represents the F1-score. On both datasets, optimal performance is achieved with a dictionary size of 200. The fusion effects are related to the fusion ratio of in-sample and cross-sample knowledge selections. Our experiments demonstrate that achieving an appropriate fusion of in-sample and cross-sample knowledge selections leads to optimal results.

4.8. Visualization

To further illustrate how CATDM and CCoAt enhance the modal fusion, we visualize the fusion results for four different models on the CMU-MOSI testing set in Fig. 7. In addition, we have visualized the results on the CMU-MOSEI* testing set under the OOD setting in Fig. 8. Specifically, we visualize the base model without CATDM and CCoAt in Figs. 7(a) and 8(a), the model with only CATDM in Figs. 7(b) and 8(b), the model with only CCoAt in Figs. 7(c) and 8(c), and the complete AtCAF model that includes both CATDM and CCoAt in Fig. 7(d) and Fig. 8(d).

Under the normal setting in Fig. 7, the base model emulates the traditional MSA model that is susceptible to multimodal noise and spurious correlations. Consequently, samples of the same class are segregated into distinct clusters, exerting pressure on the final output layer. Many samples are labeled as positive, but their representations in the embedding space are in the cluster of negative samples. Spurious correlations significantly influence the misclassified samples, and CATDM addresses this issue by adjusting the front door. Nevertheless, the boundaries between positive and negative samples could be further refined. The border areas exhibit a mixture of positive, negative, and neutral sentiments, indicating instability in their sentimental clues during fusion. In contrast, employing only CCoAt yields performance similar to the base model, suggesting bias in its fusion due to the presence of the back door. The complete AtCAF model harnesses causality-aware modal fusion, resulting in clearer boundaries between positive and negative samples compared with the CATDM-Only model, with the accurate distribution of neutral sentiments along this boundary.

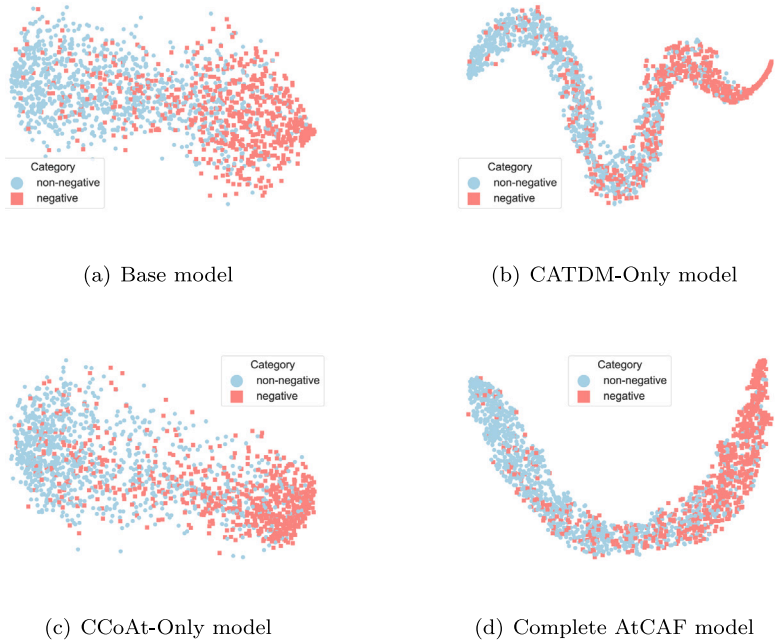
Under the OOD setting in Fig. 8, we find that although the base model roughly divided the samples into two clusters, there are two issues: 1. Samples belonging to the same class do not tightly aggregate together. 2. There is a significant mixture of samples from both classes at the geometric center of the embedding space. The first issue stems from the OOD setting, where the distribution of text across different categories is more uneven, and many samples, due to spurious correlations, are misclassified into another cluster, leading to a lack of tight aggregation within specific clusters. The second issue arises when the model is unable to extract sufficient sentimental cues from multimodal information to discern the samples; influenced by empirical risk minimization, it tends to predict a more neutral sentiment, which manifests as a convergence of a large number of samples from different classes at the geometric center of the embedding space. CATDM has alleviated the first issue by using a front door adjustment to reclassify misclassified samples due to spurious correlations, resulting in tighter intra-cluster cohesion. However, the overlap between different clusters is still significant, indicating that sentimental cues are still not easily discernible. The CCoAt module, by employing context filtering, has uncovered sentimental cues that were previously difficult to extract, reducing the number of samples located at the geometric centers of different embedding spaces, thus alleviating the second issue. In addition, we find that the counterfactual reasoning of CCoAt can also make the center of different clusters more tightly packed, thereby improving the accuracy of predictions. However, there are still many discrete points in the feature space. By combining CATDM and CCoAt, we successfully alleviated issues one and two. On one hand, the distribution of samples in different clusters has become more compact. On the other hand, the number of uncertain samples in the center of the feature space has been significantly reduced.



(a) Base model (b) CATDM-Only model

(c) CCoAt-Only model (d) Complete AtCAF model

Fig. 7. Visualization of multimodal fusion results in the embedded space using t-SNE. The blue circles represent positive samples, the red squares represent negative samples, and the green denotes neutral samples.



(a) Base model (b) CATDM-Only model

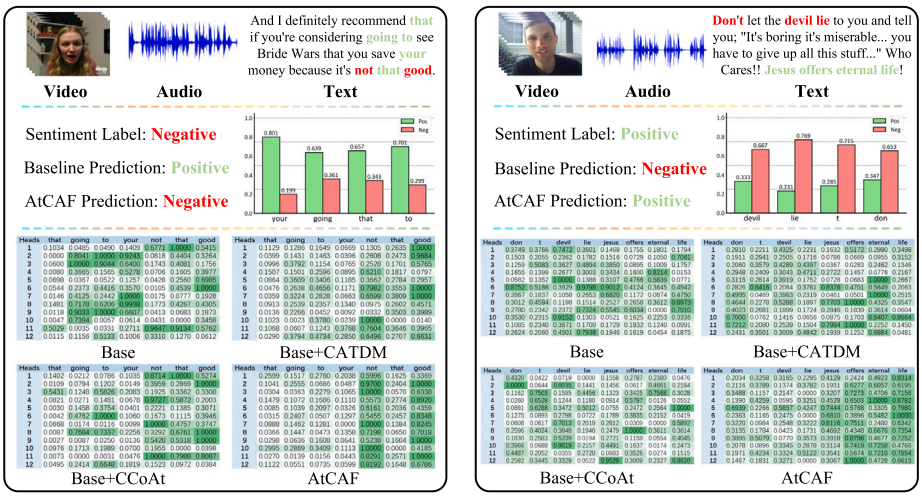
(c) CCoAt-Only model (d) Complete AtCAF model

Fig. 8. Visualization of multimodal fusion results in the embedded space using t-SNE on CMU-MOSEI* under the OOD setting. The light blue circles represent non-negative samples, and the red squares represent negative samples.

In summary, CATDM optimizes the multimodal representations of misclassified samples due to spurious correlations through the front door adjustment, making the distances within clusters more tightly packed. CCoAt, through counterfactual reasoning of the sentimental cues in the context, corrects the classification of samples located at the geometric center of the embedding space, which is challenging to discern in terms of sentimental information, significantly reducing the number of samples at the cluster boundary areas. Moreover, when combining CATDM and CCoAt, they both play their respective roles effectively, learning optimal multimodal representations in both normal and OOD settings. In real-world scenarios, multimodal data imbalance is widespread, particularly evident in the form of imbalanced data labels and the challenge of semantic extraction from samples. The aforementioned experiments demonstrate AtCAF's flexibility, sustaining performance under both normal and OOD settings and making it suitable for fields such as sentiment analysis in social media posts, opinion mining in customer reviews, and emotion recognition in multimedia content.

4.9. Case study

To better analyze the role of AtCAF in capturing causal clues in multimodal sentiment analysis, we randomly selected two test samples from CMU-MOSEI for the case study. Furthermore, we compare with MMIM [12] as a normal multimodal sentiment analysis model for



(a) Case1 (b) Case2

Fig. 9. Two samples from the CMU-MOSEI test set and the comparison of AtCAF and the baseline (*i.e.* MMIM [12]). The bar charts show the proportion of tokens in the training set that are distributed among positive/negative categories(based on the BERT tokenizer). The lower part of the image displays the visualization of attention weights for each head in the second-to-last layer of BERT [56–58], which reflects the attention that each token pays to the classification head [CLS]. We perform min–max normalization for each head to enhance the distinctiveness of different tokens, with darker colors indicating greater attention.

further exploration. To further reveal the model's mechanism for alleviating spurious correlation, we provide visualizations of the attention weights for each head in the second-to-last layer of the text encoder for different models [56–58]. "Base" refers to the basic model without CATDM and CCoAt, "Base+CATDM" indicates the variant with CATDM added to the basic model, "Base+CCoAt" indicates the variant with CCoAt added to the basic model, and "AtCAF" represents the complete model.

In Fig. 9(a), MMIM erroneously classifies the sample as "positive" due to the presence of tokens leaning towards positive sentiment in the training set, such as "your", "going", "that", and "to". The model adheres to the maximum likelihood principle, wherein even non-sentimental clue tokens contribute to the prediction. However, relying solely on maximum likelihood without considering context hampers the utilization of other clues from different modalities. AtCAF integrates causal inference methods to mitigate the influence of spurious correlations, enabling accurate predictions. Thus, AtCAF discerns the narrator's emphatic tone and affirmative expression when processing "not that good", affirming that "not good" is the central viewpoint of this video clip. Through the attention maps, we can observe that both the "Base" and "Base+CCoAt" models tend to focus on spurious correlations even with the guidance of other modalities, while the "Base+CATDM" variant mitigates this issue but some attention heads still struggle to focus on the key words of the text. By combining CATDM and CCoAt, the complete "AtCAF" model not only alleviates spurious correlations but also encourages different attention heads to be more focused on representing key words from different views, leading to more consistent sentimental judgments. Different modules work as expected when processing biased inputs.

In Fig. 9(b), the complex text contains numerous tokens associated with negativity from the training set, including both non-sentimental clue tokens like "devil", "don", "lie", and "t", and sentimental tokens such as "boring" and "miserable". Like many multimodal sentiment analysis models, MMIM misclassifies sentiment polarity due to its over reliance on the text and misinterpreting the smile in the visual as numbness. In contrast, AtCAF discerns the calm tone of the speaker, the determined gaze, and the confident smile. Leveraging multimodal sentimental clues, AtCAF accurately interprets the text as expressing optimism and positivity. Furthermore, through attention visualization, case 2 confirms the conclusions we reached about the debiasing mechanism in case 1, where CATDM calibrates attention to the key words for sentimental analysis, and CCoAt promotes the attention concentration of key words across different attention heads. In both cases, AtCAF produces the best text representation results.

As is widely known, smiles do not always indicate happiness, and sadness does not always result in tears. These emotions are expressed to varying degrees within multimodal sentimental clues. The method of aggregating information can significantly impact the interpretation of sentiment, even for identical text, speech, and video inputs. The aforementioned analysis underscores AtCAF's ability to discern effective information within multimodal contexts, thereby achieving robust tracing of sentimental causality-aware capability lacking in traditional multimodal sentiment analysis models.

Table 11
Comparison of the parameter count and FLOPs between the SOTA methods and AtCAF.

Methods	Params/M	FLOPs/G	MOSI-ACC2	MOSI-F1
MAG-BERT	86.870	4.312	–/83.5	–/83.5
TETFN	86.612	4.298	84.05/86.10	83.83/86.07
AOBERT	344.567	66.934	85.2/85.6	85.4/86.4
AtCAF	**86.233**	**4.296**	**87.03/88.57**	**86.96/88.53**

4.10. Model complexity

To further analyze the improvements in AtCAF beyond accuracy, we compare its parameters and FLOPs with those of three other SOTA methods, as shown in Table 11. It is worth mentioning that AtCAF significantly outperforms all the sota methods in accuracy and has a smaller model complexity. The results demonstrate that our model significantly outperforms other baselines on the CMU-MOSI dataset, even with a relatively smaller parameter count and computational load. Specifically, it exceeds MAG-BERT [25] and TETFN [7] by an average of 2.8% when compared with models of similar computational volume, and it surpasses the more complex AOBERT [13] by an average of 2.1%. This confirms that our proposed AtCAF balances accuracy and model complexity, which is suitable for practical applications.

5. Conclusion

We introduce the **A**ttention-based **C**ausality-**A**ware **F**usion network (AtCAF) for multimodal sentiment analysis. Our model consists of

two core modules: CATDM, which is used to capture causality-aware text representations to eliminate spurious correlations in the dataset, and CCoAt, which is used to filter out irrelevant contexts. AtCAF has achieved state-of-the-art results on public datasets in both normal and out-of-distribution settings. AtCAF is the first work to discover causal relationships for both unimodal representation and multimodal fusion and construct a complete causal effect chain to fully convey the causal path from user input and model prediction. AtCAF demonstrates strong adaptability to the widespread issues of imbalanced data and complex contexts in the real world, ensuring fairer and more substantiated outcomes in sensitive application scenarios such as social media analytics, e-commerce analytics, mental health assessment, etc. Future research can explore multimodal sentiment analysis from an interpretability perspective derived from causal inference. Additionally, there is potential for designing lighter-weight models grounded in causal inference principles.

CRediT authorship contribution statement

Changqin Huang: Writing – review & editing, Validation, Methodology, Conceptualization, Formal analysis, Investigation, Software. **Jili Chen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization, Formal analysis, Investigation. **Qionghao Huang:** Writing – review & editing, Visualization, Validation, Methodology. **Shijin Wang:** Writing – review & editing, Validation, Investigation, Visualization. **Yaxin Tu:** Investigation, Validation, Visualization, Writing – review & editing. **Xiaodi Huang:** Writing – review & editing, Validation, Methodology, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (No. 2022C03106), in part by the National Natural Science Foundation of China (No. 62337001, 62207028), and the Zhejiang Provincial Natural Science Foundation (No. LY23F020009), and Special Research Project of Zhejiang Normal University on Serving Provincial Strategic Planning and Promoting the Construction of Common Prosperity.

References

[1] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.

[2] R. Das, T.D. Singh, Multimodal sentiment analysis: A survey of methods, trends and challenges, ACM Comput. Surv. (2023).

[3] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.

[4] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A.B. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2247–2256.

[5] P. Koromilas, M.A. Nicolaou, T. Giannakopoulos, Y. Panagakis, MMATR: A lightweight approach for multimodal sentiment analysis based on tensor methods, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.

[6] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2019, NIH Public Access, 2019, p. 6558.

[7] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognit. 136 (2023) 109259.

[8] C. Huang, J. Zhang, X. Wu, Y. Wang, M. Li, X. Huang, TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis, Knowl.-Based Syst. 269 (2023) 110502.

[9] H. Cheng, Z. Yang, X. Zhang, Y. Yang, Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion, IEEE Trans. Affect. Comput. (2023).

[10] Z. Li, Q. Guo, Y. Pan, W. Ding, J. Yu, Y. Zhang, W. Liu, H. Chen, H. Wang, Y. Xie, Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis, Inf. Fusion 99 (2023) 101891.

[11] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 10790–10797.

[12] W. Han, H. Chen, S. Poria, 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, 2021, pp. 9180–9192.

[13] K. Kim, S. Park, AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis, Inf. Fusion 92 (2023) 37–45.

[14] T. Wang, J. Huang, H. Zhang, Q. Sun, Visual commonsense r-cnn, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10760–10770.

[15] T. Wang, C. Zhou, Q. Sun, H. Zhang, Causal attention for unbiased visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3091–3100.

[16] X. Yang, H. Zhang, G. Qi, J. Cai, Causal attention for vision-language tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9847–9857.

[17] X. Yang, H. Zhang, J. Cai, Deconfounded image captioning: A causal retrospect, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[18] J. Pearl, Causality, Cambridge University Press, 2009.

[19] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88.

[20] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.

[21] M.K. Hasan, W. Rahman, A. Zadeh, J. Zhong, M.I. Tanveer, L.-P. Morency, et al., UR-FUNNY: A multimodal language dataset for understanding humor, 2019, arXiv preprint arXiv:1904.06618.

[22] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[23] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[25] W. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2020, NIH Public Access, 2020, p. 2359.

[26] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8992–8999.

[27] D. Janzing, B. Schölkopf, Causal inference using the algorithmic Markov condition, IEEE Trans. Inform. Theory 56 (10) (2010) 5168–5194.

[28] P. Kamath, A. Tangella, D. Sutherland, N. Srebro, Does invariant risk minimization capture invariance? in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 4069–4077.

[29] S. Seo, J.-Y. Lee, B. Han, Information-theoretic bias reduction via causal view of spurious correlation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2180–2188.

[30] Y. Rao, G. Chen, J. Lu, J. Zhou, Counterfactual attention learning for fine-grained visual categorization and re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1025–1034.

[31] Y. Liu, G. Li, L. Lin, Cross-modal causal relational reasoning for event-level visual question answering, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[32] P.-J. Huang, H. Xie, H.-C. Huang, H.-H. Shuai, W.-H. Cheng, CA-FER: Mitigating spurious correlation with counterfactual attention in facial expression recognition, IEEE Trans. Affect. Comput. (2023).

[33] T. Sun, W. Wang, L. Jing, Y. Cui, X. Song, L. Nie, Counterfactual reasoning for out-of-distribution multimodal sentiment analysis, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 15–23.
[34] T. Sun, J. Ni, W. Wang, L. Jing, Y. Wei, L. Nie, General debiasing for multimodal sentiment analysis, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5861–5869.
[35] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan kaufmann, 1988.
[36] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP—A collaborative voice analysis repository for speech technologies, in: 2014 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp), IEEE, 2014, pp. 960–964.
[37] T.B. Moeslund, A. Hilton, V. Krüger, L. Sigal, Visual Analysis of Humans, Springer, 2011.
[38] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[39] J. Pearl, M. Glymour, N.P. Jewell, Causal Inference in Statistics: A Primer, John Wiley & Sons, 2016.
[40] C. Huang, H. Wei, Q. Huang, F. Jiang, Z. Han, X. Huang, Learning consistent representations with temporal and causal enhancement for knowledge tracing, Expert Syst. Appl. 245 (2024) 123128.
[41] Z. Chen, L. Hu, W. Li, Y. Shao, L. Nie, Causal intervention and counterfactual reasoning for multi-modal fake news detection, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 627–638.
[42] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, PMLR, 2015, pp. 2048–2057.
[43] J. Pearl, Direct and indirect effects, in: Probabilistic and Causal Inference: The Works of Judea Pearl, 2022, pp. 373–392.
[44] T. VanderWeele, Explanation in Causal Inference: Methods for Mediation and Interaction, Oxford University Press, 2015.
[45] Y. Hagmayer, S.A. Sloman, D.A. Lagnado, M.R. Waldmann, Causal reasoning through intervention, Causal Learn.: Psychol. Philos. Comput. (2007) 86–100.
[46] E.H. Aarts, et al., Simulated Annealing: Theory and Applications, Reidel, 1987.
[47] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 6–15.
[48] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022).
[49] H. Sun, H. Wang, J. Liu, Y.-W. Chen, L. Lin, CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3722–3729.
[50] R. Lin, H. Hu, Multi-task momentum distillation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2023).
[51] H. Shi, Y. Pu, Z. Zhao, J. Huang, D. Zhou, D. Xu, J. Cao, Co-space representation interaction network for multimodal sentiment analysis, Knowl.-Based Syst. 283 (2024) 111149.
[52] J. Huang, J. Zhou, Z. Tang, J. Lin, C.Y.-C. Chen, TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis, Knowl.-Based Syst. 285 (2024) 111346.
[53] C. Gan, Y. Tang, X. Fu, Q. Zhu, D.K. Jain, S. García, Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation, Knowl.-Based Syst. (2024) 111982.
[54] C. Fan, K. Zhu, J. Tao, G. Yi, J. Xue, Z. Lv, Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2024).
[55] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Adv. Neural Inf. Process. Syst. 32 (2019).
[56] H. Yang, Bert meets chinese word segmentation, 2019, arXiv preprint arXiv: 1909.09292.
[57] J. Yang, M. Wang, H. Zhou, C. Zhao, W. Zhang, Y. Yu, L. Li, Towards making the most of bert in neural machine translation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 9378–9385.
[58] K. Lu, Z. Wang, P. Mardziel, A. Datta, Influence patterns for explaining information flow in bert, Adv. Neural Inf. Process. Syst. 34 (2021) 4461–4474.