Complex & Intelligent Systems (2025) 11:196
https://doi.org/10.1007/s40747-025-01806-y

ORIGINAL ARTICLE

# H$^2$CAN: heterogeneous hypergraph attention network with counterfactual learning for multimodal sentiment analysis

Changqin Huang[1,2] · Zhenheng Lin[1,3] · Qionghao Huang[1,2] · Xiaodi Huang[4] · Fan Jiang[1,3] · Jili Chen[1,2]

Received: 8 July 2024 / Accepted: 5 January 2025 / Published online: 28 February 2025


## Abstract

Multimodal sentiment analysis (MSA) has garnered significant attention for its immense potential in human-computer interaction. While cross-modality attention mechanisms are widely used in MSA to capture inter-modality interactions, existing methods are limited to pairwise interactions between two modalities. Additionally, these methods can not utilize the causal relationship to guide attention learning, making them susceptible to bias information. To address these limitations, we introduce a novel method called Heterogeneous Hypergraph Attention Network with Counterfactual Learning (H$^2$CAN). The method constructs a heterogeneous hypergraph based on sentiment expression characteristics and employs Heterogeneous Hypergraph Attention Networks (HHGAT) to capture interactions beyond pairwise constraints. Furthermore, it mitigates the effects of bias through a Counterfactual Intervention Task (CIT). Our model comprises two main branches: hypergraph fusion and counterfactual fusion. The former uses HHGAT to capture inter-modality interactions, while the latter constructs a counterfactual world using Gaussian distribution and additional weighting for the biased modality. The CIT leverages causal inference to maximize the prediction discrepancy between the two branches, guiding attention learning in the hypergraph fusion branch. We utilize unimodal labels to help the model adaptively identify the biased modality, thereby enhancing the handling of bias information. Experiments on three mainstream datasets demonstrate that H$^2$CAN sets a new benchmark.

**Keywords** Multimodal sentiment analysis · Modality interaction · Heterogeneous hypergraph · Counterfactual learning

## Introduction

In the digital communication era, the mediums for expressing sentiments have evolved from simple text messages to complex videos, presenting both challenges and opportunities for sentiment analysis. Relying solely on textual semantics in multimodal data, such as videos, may result in inaccuracies. Thus, incorporating nonverbal modalities is essential to counter the emotional biases inherent in single-modality sentiment analysis [1]. For example, a sarcastic remark in a video might be incorrectly interpreted as positive if only the textual content is considered. It is essential to consider both the speaker's tone of voice and facial expressions simultaneously to correct such misinterpretations. Consequently, multimodal sentiment analysis (MSA) has gained popularity among researchers as it achieves more accurate sentiment

author wants to mention Dr. Fan Jiang's and Prof. Changqin Huang equally contributed

✉ Changqin Huang
cqhuang@zju.edu.cn

✉ Fan Jiang
fanjiang@zjnu.edu.cn

Zhenheng Lin
a83531777@zjnu.edu.cn

Qionghao Huang
qhhuang@m.scnu.edu.cn

Xiaodi Huang
xhuang@csu.edu.au

Jili Chen
irelia@zjnu.edu.cn

[1] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321004, Zhejiang, China

[2] School of Education, Zhejiang Normal University, Jinhua 321004, Zhejiang, China

[3] School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, Zhejiang, China

[4] School of Computing, Mathematics and Engineering, Charles Sturt University, Albury, NSW 2640, Australia

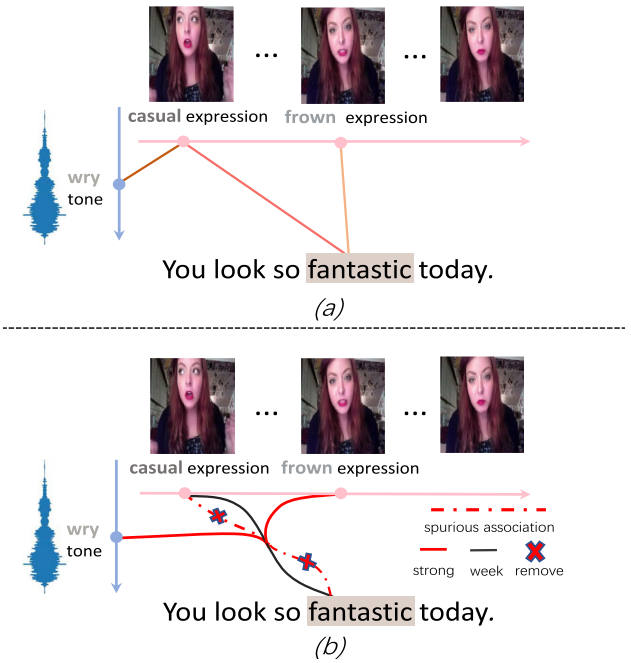


**Fig. 1** The example case illustrates the advantages of our model. **a** The scope of interactions captured by the attention methods in general MSA models is limited to two nodes within two modalities. **b** Our model enables the modeling of relationships between multiple nodes across multiple modalities. Furthermore, our model can mitigate the effects of bias information

understanding by leveraging complementary information from text modality and nonverbal modalities (i.e., visual and audio) [2].

The main challenge of multimodal sentiment analysis is to effectively capture inter-modality interaction information [3]. Early MSA studies attempt to solve this challenge through tensor outer product or LSTM [4–6], yet these methods failed to capture fine-grained interaction information. Attention mechanisms have been explored [7] to address this limitation. MulT [8] introduces transformers into the MSA task, employing cross-modality attention to capture inter-modality interactions. On this basis, TeTFN [9] proposes a text-centered cross-modality attention mechanism. Besides, there are studies [10, 11] that introduce graph-based methods to overcome the challenges of MSA. These methods consider the time steps of multimodal sequence data as nodes to construct a graph, then apply graph attention networks to explore interactions between modalities. These cross-modality attention-based models yield promising results; however, there still exist two issues.

The scope of interactions captured by cross-modality attention methods within existing MSA models is limited to just two time steps (nodes) within two modalities. Because interactions among different modalities in emotional expression are more than dyadic (pairwise), this constraint hinders the full exploitation of inter-modality interaction information [12]. The case depicted in Fig. 1 expresses a satirical, negative view. As shown in Fig. 1a, the pairwise interaction between the word 'fantastic' in the text modality and the 'casual expression' in the visual modality conveys a positive sentiment, which is contrary to the actual negative sentiment of the case. Sentiment polarity is corrected when incorporating the 'wry tone' and the 'frown expression' (as depicted in Fig. 1b) to model beyond-pairwise interactions across multiple modalities. Given the limitations of current attention methods, a crucial issue arises: **(Q1) How can we model interactions beyond pairwise among multiple time steps across various modalities using an attention method?**

General MSA models are typically supervised solely by the final task loss, lacking a powerful supervisory signal to guide attention learning [13]. This weakly-supervised approach ignores causal relationships between attention and results, making attention methods prone to learning spurious associations with bias information [14]. As shown in Fig. 1b, 'casual expressions' and the word 'fantastic', which occur more frequently in neutral and positive samples respectively, may lead attention models to use them as discriminative cues. When such bias information receives much attention, as depicted by the red dashed line (i.e., spurious associations), it may mislead the model. Moreover, the impact of bias information from different modalities on the model varies in magnitude [15]. For instance, the word 'fantastic' has a greater effect than 'casual expression' because the former exhibits a larger sentiment gap relative to the actual sentiment of the case. This raises another issue: **(Q2) How can we mitigate the effects of bias information from different modalities by guiding attention learning?**

To address the above two issues, we present a Heterogeneous Hypergraph Attention Network with Counterfactual Learning ($H^2$CAN). To address **(Q1)**, our approach first introduces heterogeneous hypergraphs into MSA to overcome the limitation of pairwise interaction between two modalities. Specifically, we consider the temporal relations of nonverbal modality and the syntax-aware relations of text modality [16, 17]. To capture these relations, we construct a heterogeneous hypergraph based on temporal links and the dependency tree of the sentence. Next, we construct a hypergraph fusion branch based on the heterogeneous hypergraph and utilize a heterogeneous hypergraph attention network to capture beyond-pairwise relations across different modalities.

To address **(Q2)**, our approach aims to guide attention learning by focusing on the underlying causal relationships between modalities, thereby mitigating the effects of bias information. We introduce a Counterfactual Intervention Task (CIT) based on sentiment polarity classification to achieve this. First, we create a counterfactual intervention branch derived from the hypergraph fusion branch. Given

that bias information from different modalities has varying impacts, we identify the modality with the largest discrepancy between unimodal and multimodal labels as the biased modality, as this modality's bias information has a more significant effect. To generate unimodal labels for adaptively identifying biased modalities, we incorporate a Unimodal Label Generation Module (ULGM) [18]. The popular counterfactual approaches are based on the normal distribution, but they ignore the differences between different modalities under multimodal situations. Therefore, in the counterfactual intervention branch, we generate attention scores using a Gaussian distribution and assign additional weights to the attention scores corresponding to the biased modality to magnify the effects of bias information. Finally, CIT employs causal inference to direct the model's focus on the causal relationship between attention and prediction by maximizing the difference between the predictions of the two branches, thereby mitigating the effects of bias information.

In summary, our work makes the following contributions:

- We introduce heterogeneous hypergraphs to MSA and propose a novel model named $H^2$CAN, which can model beyond-pairwise interactions among multiple nodes across modalities.
- We design a heterogeneous hypergraph construction method based on the temporal relations of nonverbal modalities and the syntax-aware relations of text modalities, enabling the modeling of inter-modality interactions.
- We introduce counterfactual learning and devise a Counterfactual Intervention Task that enables the model to employ causal inference to mitigate the effects of bias information from multiple modalities, thereby improving the performance of the model.
- In a series of evaluations on benchmark datasets, $H^2$CAN surpasses several competitive models on the MSA task.

## Related work

### Multimodal sentiment analysis

In sentiment analysis, in addition to the semantic information in text, emotional cues from audio and visual modalities are also essential [19, 20]. As an emerging area of affective computing, multimodal sentiment analysis integrates information from text, images, and audio to assess human sentiment in video clips. Previous studies relied on straightforward methods, such as concatenation and summation, to obtain interaction information across modalities. Yang et al. [21] develop a novel joint contrastive learning framework that aims to dissect similarity and dissimilarity features through contrast learning and unimodal tasks and finally concatenate these features together for final prediction. MISA [22] maps unimodal features into both modality-specific and modality-invariant subspaces, effectively capturing unique and shared information across modalities.

To capture inter-modality interactions more effectively, MulT [8] devise a cross-modality attention method to investigate fine-grained interactions between different pairs of time steps in multimodal sequences. Huang et al. [23] argue that text contains richer emotional information and introduces a modality fusion paradigm centered on the text modality. TMBL [3] introduces CLS and PE feature vectors into cross-modality transformers to retain invariant and specific information across modalities. PEST [24] mitigates heterogeneity in multimodal fusion by mapping features from diverse modalities into a unified feature space. CMHFM [25] enhances multimodal fusion through the integration of multimodal combination subtasks. With the rise in popularity of graph neural networks (GCN), Yang et al. [26] first introduce GCN to multimodal sentiment analysis. MDH [11] models intra-modality interactions through unimodal hypergraphs and captures inter-modality interactions through attention mechanisms.

However, these models fail to explore beyond-pairwise interactions between multiple modalities, as the scope of interactions captured by their attention methods is limited to two time steps within two modalities. Furthermore, the attention methods used in these models are supervised only by the final predictive loss, thus making them susceptible to bias information from different modalities. In contrast, $H^2$CAN not only models beyond-pairwise interactions between multiple nodes across multiple modalities but also utilizes causal inference to guide attention learning.

### Hypergraph network

Hypergraph networks have attracted widespread interest for their capability to capture complex information among multiple nodes via hyperedges [27]. In multimodal learning, HANs [28] utilize random walks to construct hypergraphs for image and text modalities, creating associations between them through a co-attention map. VM-HAN [29] constructs hypergraphs based on node similarity to capture higher-order relationships among nodes. HHGSA [30] constructs hypergraphs based on the associative properties of object nodes to learn complex relationships among different objects. Nevertheless, the hypergraph construction methods employed in these models are not well-suited for the MSA task, which suffers from modality heterogeneity [31]. Additionally, these methods overlook the importance of guiding attention learning within the hypergraph network. In response, we propose a new heterogeneous hypergraph construction method for modeling emotional interactions across text, visual, and audio modalities, and introduce the Counterfactual Intervention Task to guide hypergraph attention learning.

### Causal inference

Causal inference is extensively applied for deep learning domains to improve the performance and interpretability of models, such as recommendation systems, information retrieval, computer vision, etc. CAL [14] constructs a random attention graph to assist the model in learning more effective attention. MMRec [32] utilize user-uninteracted items to purify user preference-relevant information. CausVSR [33] establishes emotional causality based on a sequence of emotional perceptions to counter challenges from confounders. In the field of multimodal sentiment analysis, Sun et al. [34] employ multimodal cues to mitigate spurious correlations from text. AtCAF [35] captures causality-aware text representations before modality fusion through front-door adjustment. While these works have made significant advancements, they mainly address bias information from the text modality. However, in complex multimodal data, bias information from various modalities can influence the model at different magnitudes. In response to this, we propose a Counterfactual Intervention Task to guide attention learning, effectively reducing the impact of bias information across all modalities, especially that with the greatest influence.

## $H^2$CAN

This section begins with a description of MSA task. Following that, we will detail the framework of $H^2$CAN. Figure 2 provides the workflow of $H^2$CAN: unimodal feature encoding, counterfactual multimodal fusion, and prediction and optimization.

**Task description**. A video containing text, visual, and audio modalities is segmented into multiple non-overlapping clips. The MSA task aims to predict an emotional intensity score y for each clip, ranging from $-3.0$ to 3.0. Furthermore, sentiment polarities can be identified based on emotional intensity scores, where positive values indicate positive sentiments, negative values indicate negative sentiments and zero values indicate neutral sentiments.

### Unimodal feature encode

Suppose the model accepts as input a multimodal sequence $u = u_t, u_v, u_a$, where the sequence consists of $l$ time steps. Here, $t$ stands for text, $v$ for visual, and $a$ for acoustic modalities, respectively. Firstly, we use modality-specific encoders to extract initial representations of these multimodal sequences. We use pre-trained BERT [36] to encode the text modality sequence $u_t$, thus obtaining tokens $x_t$ containing semantic information:

$$e_t = BERT(u_t; \theta^{BERT}) \in \mathbb{R}^{l \times d_t}, \tag{1}$$

where $\theta^{BERT}$ denotes the learnable parameter of BERT, $d_t$ denotes the features dimension, and $l$ indicates the length of text modality sequence.

For the nonverbal modalities, we extract a set of low-level visual representations and audio representations that can reflect emotional information from raw data using Facet [37] and COVAREP [38], respectively. Then, we employ modality-specific sequence encoders to model the context of the nonverbal modalities. Given that LSTM [39] shows comparable performance to the Transformer [40] on the MSA task (as shown in Table 2) with fewer parameters, and its use in prior work [9, 41], we adopt LSTM as our sequence encoder. The nonverbal features $e_v$ and $e_a$ are computed as follows:

$$e_v = Bi\text{-}LSTM_v(u_v; \theta_v^{LSTM}) \in \mathbb{R}^{l \times d_v}, \tag{2}$$

$$e_a = Bi\text{-}LSTM_a(u_a; \theta_a^{LSTM}) \in \mathbb{R}^{l \times d_a}. \tag{3}$$

Finally, these features are mapped into feature spaces with the same dimensions through three feedforward layers, resulting in three feature representations $x_t$, $x_a$, and $x_v$ with the same shapes:

$$x_n = FFN_n(e_n; \theta_n^{FFN}) \in \mathbb{R}^{l \times d}, \quad n \in \{a, v, t\}, \tag{4}$$

where $FFN(\cdot)$ represents a 1D temporal convolution layer.

### Counterfactual multimodal fusion

In the counterfactual multimodal fusion phase, the heterogeneous hypergraph attention network (HHGAT) is utilized to capture inter-modal beyond-pairwise interactions. As shown in Fig. 2, the hypergraph fusion branch integrates interaction information between multiple nodes across multiple modalities. Then, we construct the counterfactual fusion branch based on the hypergraph fusion branch. Specifically, we leverage the multimodal representation from the hypergraph fusion branch to identify the biased modality and construct a counterfactual world by assigning extra weights to it in the counterfactual fusion branch. Finally, the Counterfactual Intervention Task (CIT) is performed to mitigate the effects of bias in the hypergraph fusion branch.

**Heterogeneous hypergraph construction**. The most commonly used methods for Hypergraph construction, such as KNN and K-means, are computationally expensive and unsuitable for heterogeneous data. Therefore, we propose a novel static heterogeneous hypergraph construction method that requires no computational resources. Specifically, we consider time steps of multimodal data as nodes and construct

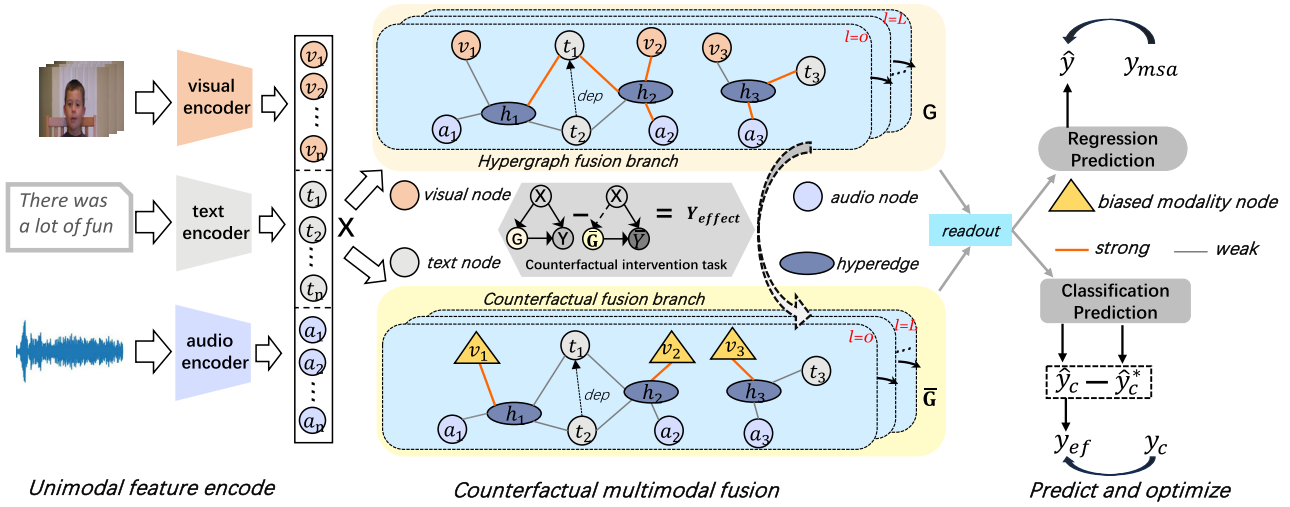


**Fig. 2** The architecture of our H$^2$CAN. The dashed arrow from the hypergraph fusion branch to the counterfactual fusion branch indicates that the latter builds upon the former. Within the heterogeneous hypergraph, dashed lines between text nodes indicate dependencies between words. ULGM is omitted in the figure

hyperedges based on a text node. For each text node $x_{ti}$, we select $s$ audio nodes and visual nodes that are temporally closest to that text node. The number of nodes $s$ of the non-verbal modality is obtained by dividing the sequence length $l$ by the hyperparameter $k$. In addition, to integrate semantic information from text modality, we employ the spaCy toolkit to construct a dependency tree based on the sentence structure. Text nodes exhibiting a dependency relationship with the node $x_{ti}$ are incorporated into this hyperedge $\mathcal{H}_i$.

$$\mathcal{H}_i = \{x_{ti}\} \cup N_s^A(x_{ti}) \cup N_s^V(x_{ti}) \cup N_d(x_{ti}),$$
$$s = \left\lfloor \frac{l}{k} \right\rfloor, \tag{5}$$

where $N_s^A(x_{ti})$ and $N_s^V(x_{ti})$ denote the sets of audio and visual nodes, respectively, that are closest to $x_{ti}$ in terms of temporal proximity. $N_d(x_{ti})$ represents the set of text nodes that have a dependency relationship with text node $x_{ti}$, as determined by the dependency tree constructed from the sentence. $\lfloor \cdot \rfloor$ indicates rounding down. The variable $l$ indicates the length of text modality sequence. Thus, we obtain a heterogeneous hypergraph with l hyperedges.

**Modal fusion**. Based on the graph derived from heterogeneous hypergraph construction, we employ the Heterogeneous Hypergraph Attention Network (HHGAT) to model inter-modality interactions. First, we aggregate information from the nodes within hyperedges. Specifically, for each hyperedge $i$ in an HHGAT layer, we concatenate representations of the current hyperedge $i$ and the node $j$ within this hyperedge and map them into an attention scalar $\hat{b}$. Then, the attention scalars of all nodes within the hyperedge are normalized by softmax. The representations of hyperedge $i$ are updated by taking a weighted.

$$\hat{b}_{ij}^{(I)} = LeakyReLU(\check{b}^{(I)}[W_e^{(I)} e_i^{(I-1)} || W_n^{(I)} x_j^{(I-1)}]),$$
$$n \in \{a, v, t\}, \tag{6}$$

$$c_{ij}^{(I)} = softmax_j(\hat{b}_{ij}^{(I)}) = \frac{\hat{b}_{ij}^{(I)}}{\sum_{u \in \mathcal{H}_i} \hat{b}_{iu}^{(I)}}, \tag{7}$$

$$e_i^{(I)} = \sum_{u \in \mathcal{H}_i} x_{iu}^{(I-1)} c_{iu}^{(I)}, \tag{8}$$

where $W^{(I)}$ is the parameter matrix of HHGAT at $I$th layer and $n$ is chosen according to the modality of node $x_j$. Additionally, $\check{b}$ denotes a parameter vector, and $||$ denotes the concatenation operation. Notably, the initial representation $e_i^{(0)}$ of hyperedge is obtained by averaging the nodes within the hyperedge. $LeakyReLu(\cdot)$ is an activation function.

After learning the hyperedge representations, we update the node representations by diffusing the information of hyperedges. Specifically, for each node $i$ in an HHGAT layer, we concatenate the representations of the current node $i$ and the hyperedge $j$ that includes node $i$ and map them into an attention scalar. Then, we normalize the attention scalars of all hyperedges that include node $i$ using the softmax function. The embedding of node $i$ gets updated by taking a weighted aggregation of the embeddings of hyperedges that include node $i$.

$$\bar{b}_{ij}^{(I)} = LeakyReLU(\tilde{b}^{(I)}[\hat{W}_n^{(I)} x_i^{(I-1)} || \hat{W}_h^{(I)} h_j^{(I-1)}]),$$
$$n \in \{a, v, t\}, \tag{9}$$

$$\hat{c}_{ij}^{(I)} = softmax_j(\bar{b}_{ij}^{(I)}) = \frac{\bar{b}_{ij}^{(I)}}{\sum_{u\in\mathcal{D}_i}\bar{b}_{iu}^{(I)}}, \tag{10}$$

$$x_i^{(I)} = \sum_{u\in\mathcal{D}_i} e_{iu}^{(I)}\hat{c}_{iu}^{(I)}, \tag{11}$$

where $\hat{W}^{(I)}$ denotes the parameter matrix responsible for hyperedge diffusion in HHGAT at the $I$th layer, and $\tilde{b}$ is a parameter vector. Furthermore, $\mathcal{D}_i$ denotes the set of hyperedges that include node $i$.

In each of the two fusion branches of HHGAT, we aggregate all nodes in the last layer to obtain the multimodal representation $o$ and the counterfactual multimodal representation $o^*$. The former is derived from the hypergraph fusion branch, while the latter is from the counterfactual fusion branch.

$$o = readout\Big(HHGATs(x_t, x_v, x_a)\Big), \tag{12}$$

$$o^* = readout\Big(HHGATs^*(x_t, x_v, x_a)\Big), \tag{13}$$

where $HHGATs(\cdot)$ denotes the operation of HHGAT layers and $HHGATs^*(\cdot)$ denotes the operation of HHGAT layers in the counterfactual fusion branch. Notably, the counterfactual fusion branch is utilized exclusively during the training phase. We employ $readout(\cdot)$ functions to aggregate the representations of nodes into a graph-level representation following You et al. [42].

**Identification of biased modality**. We introduce the Unimodal Label Generation Module (ULGM) to generate unimodal sentiment score labels, which are used to identify the biased modality, where bias information has a greater effect. Firstly, we generate unimodal labels by utilizing multimodal sentiment score labels, representations from the final time step of unimodal sequences, and multimodal representation. Subsequently, we calculate the difference between the multimodal and unimodal sentiment score labels. Finally, we select the modality with the largest difference from the multimodal sentiment scores as the biased modality.

$$dist_n = |y_{msa} - ULGM(y_{msa}, x_n, o)|, \tag{14}$$

$$n^* = \arg\max_n \text{dist}_n, \tag{15}$$

where $n$ belongs to the set $\{t, a, v\}$, corresponding respectively to the text, audio, and visual modalities. $y_{msa}$ denotes the multimodal sentiment score labels and $ULGM(\cdot)$ represents the unimodal label generation module. $|\cdot|$ indicates the absolute value and $n^*$ indicates the biased modality. Notably, we only need to select the biased modality during the training stage.

**Counterfactual intervention task**. We introduce the Counterfactual Intervention Task (CIT) through a causal graph as described in Fig. 2. The causal graph is depicted as a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N}$ represents the set of variables treated as nodes, and $\mathcal{E}$ represents the set of edges that depict interactions among these variables. In our example, the causal graph is composed of three nodes: $X$ (features of different modalities), $G$ (attention scores during the first stage of message passing in HHGAT), and $Y$ (prediction result). Furthermore, the edge $X \rightarrow G$ signifies that the model generates attention scores based on the features from different modalities, while the edge $(X, G) \rightarrow Y$ indicates that different modalities' features and the attention scores jointly determine the prediction result. Due to the presence of confounders (i.e., bias information) in $X$, the model can be misled, resulting in spurious associations in attention learning. Motivated by causal analysis research [43], we utilize counterfactual intervention to evaluate learned cross-modality attention.

In causal inference, we modify the values of variables and observe the resultant effects to analyze causality. This method, referred to as intervention [44], can be denoted by $do(\cdot)$. We construct a counterfactual world through the intervention $do(G = \bar{\mathbf{G}})$, where $\bar{\mathbf{G}}$ denotes a non-existent attention scores. Previous research [45] constructs counterfactual worlds based solely on random distribution. Under multimodal situations, it is necessary to consider the differences between different modalities. Therefore, we generate counterfactual attention scores using a Gaussian distribution and apply additional weights to the attention scores corresponding to the biased modality to construct the counterfactual world:

$$\bar{G}_s = \begin{cases} (1+\alpha)\hat{G}_s, & \text{if } s = n^* \\ \hat{G}_s, & \text{else} \end{cases},$$

$$s \in \{a, v, t\}, \tag{16}$$

$$\hat{y}_c = P(G = \mathbf{G}, X = \mathbf{X}) = \mathcal{C}(o), \tag{17}$$

$$\tilde{y}_c = P(do(G = \bar{\mathbf{G}}), X = \mathbf{X}) = \mathcal{C}(o^*), \tag{18}$$

where $\hat{G}_s$ corresponds to the counterfactual attention scores of modality $s$, obtained from the predefined distribution, and $\alpha$ is the hyperparameter. $\hat{y}_c$ denotes the prediction based on learned cross-modality attention, and $\tilde{y}_c$ represents the prediction based on the counterfactual world that suffers from the influence of bias information. $\mathcal{C}$ denotes the shared classification header, including a ReLU function and a feedforward layer.

According to prior research [46], we determine the direct causality effect of learned attention and the counterfactual world by calculating the difference between their prediction results.

$$y_{ef} = \hat{y}_c - \tilde{y}_c. \tag{19}$$

We utilize the direct causality effect $y_{ef}$ as a supervisory signal to optimize the model, thereby mitigating the impact of bias information in the hypergraph fusion branch.

$$\mathcal{L}_{ci} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{ce}(y_{ef}^{i}, y_{c}^{i}), \tag{20}$$

where $N$ represents the number of samples, $\mathcal{L}_{ce}$ represents the cross entropy loss, and $y_c$ indicates the sentiment polarity categorization label based on the multimodal sentiment score label $y_{msa}$. Notably, the Counterfactual Intervention Task and the counterfactual fusion branch are only used during the training stage. Besides, we reuse part of the hypergraph fusion branch output for the counterfactual fusion branch, thus only introducing light extra computation without additional parameters.

## Prediction and optimization

Besides the Counterfactual Intervention Task as an auxiliary task, we also perform a multimodal sentiment analysis task to optimize our proposed H$^2$CAN.

**Sentiment prediction**. Finally, we use the representation $o$ as input to a multilayer perceptron to predict the multimodal sentiment score.

$$\hat{y} = MLP(o; \theta_{MLP}), \tag{21}$$

where $MLP(\cdot)$ represents a multilayer perceptron consisting of a ReLU function and a feedforward layer, and $\theta_{MLP}$ denotes the parameters of this multilayer perceptron.

**Optimization Objectives**. We optimize the model by employing both the original MSA task and the Counterfactual Intervention Task. The mean absolute error is used to calculate the MSA task loss. The complete training loss is then computed as the sum of the MSA task loss and the Counterfactual Intervention Task loss.

$$\mathcal{L}_{task} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}^{i} - y_{msa}^{i}|, \tag{22}$$

$$\mathcal{L} = \mathcal{L}_{task} + \mathcal{L}_{ci}, \tag{23}$$

where $\mathcal{L}$ represents the overall training loss for our model.

# Experiments

In this section, we will detail the descriptions of datasets, metrics, baseline comparisons, ablation studies of H$^2$CAN, and results analysis.

**Algorithm 1** Heterogeneous hypergraph attention network with counterfactual learning

**Input:** Multimodal sequences $\mathcal{M} = (U_t, U_a, U_v)$
**Output:** Prediction $\hat{y}$
**for** epoch = 1 ... N **do**
  **for** minibatch $\mathcal{B} = \{(u_t, u_a, u_v)\}$ sampled from $\mathcal{M}$ **do**
    Encode unimodal features $u_m$ into $x_m$ as (1) to (4), $m \in \{t, a, v\}$;
    Construct heterogeneous hypergraph as (5);
    Aggregate node and diffusion hyperedge information as (6) to (11);
    Obtain multimodal representation $o$ as (12);
    Identify biased modality as (14), (15);
    Construct counterfactual fusion branch as (16);
    Produce counterfactual multimodal representation $o^*$ as (6) to (11), and (13);
    Use cross-entropy to compute $\mathcal{L}_{ci}$ as (17) to (20);
    Obtain prediction result $\hat{y}$ as (21);
    Compute $\mathcal{L}_{task}$ for MSA task as (22);

    Add $\mathcal{L}_{task}$ and $\mathcal{L}_{ci}$ to compute $\mathcal{L}$ as (23);
  **end for**
**end for**

**Table 1** Data distribution for benchmark datasets

| Dataset | Train | Valid | Test | All |
|---|---|---|---|---|
| CMU-MOSI | 1284 | 229 | 686 | 2199 |
| CMU-MOSEI | 16,326 | 1871 | 4659 | 22,856 |
| CH-SIMS | 1368 | 456 | 457 | 2281 |

## Datasets and evaluation metrics

Our experiments are conducted on the mainstream datasets, CMU-MOSI [47], CMU-MOSEI [48] and CH-SIMS [49]. Table 1 presents the details of the dataset split.

**MOSI**. The CMU-MOSI dataset was constructed using movie review videos sourced from YouTube. It comprises 93 videos, segmented into 2199 clips based on content. Each segment contains a label value in the interval $[-3, 3]$, indicating the sentiment intensity.

**MOSEI**. The CMU-MOSEI builds upon the CMU-MOSI, consisting of 23,453 video clips, covering a diverse range of topics. In addition to sentiment score labels, each clip also includes emotion labels such as anger and fear.

**CH-SIMS**. The CH-SIMS is a Chinese dataset for multimodal sentiment analysis, comprising 2281 video clips from various media sources. Each clip is assigned a sentiment score on a scale from $-1$ (negative) to 1 (positive), with additional sentiment scores provided for each individual modality.

Following prior research [50], we evaluate our model based on regression and categorization tasks. The definitions for specific metrics are as follows.

**Mean Absolute Error (MAE)**. The average error between the predicted and actual sentiment intensity values.

**Pearson Correlation Coefficient (Corr)**. The quantification of the linear relationship between predicted and actual sentiment intensity values.

**Binary Classification Accuracy (Acc-2)**. The accuracy rate for binary classification of emotional polarity. On the MOSI and MOSEI datasets, there are two forms: negative/non-negative (including neutral) and negative/positive (excluding neutral).

**Weighted F1-score (F1-score)**. This metric balances precision and recall in binary classification. On the MOSI and MOSEI datasets, there are two forms (refer to Acc-2).

For the categorization task, we use Acc-2 and F1-score as evaluation metrics. For the regression task, we use MAE and Corr as evaluation metrics.

## Implementation details

We utilize two tools commonly used in the MSA field, COVAREP [38] and Facet [37], to extract features from nonverbal modalities. Furthermore, we employ spaCy toolkit to build a dependency tree. The optimizer for our model is Adam [51] combined with a learning rate decay strategy. BERT and other modules are trained with different learning rates: 5e−5 for the former and 5e−4 for the latter. For the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, we set the batch sizes to 32, 64, and 32, respectively. Across all datasets, we set the LSTM layer to 1 and fix the hidden dimension in both the attention module and convolution layer to 30, with the convolution kernel size fixed at 1. Other hyperparameters are tuned within the following ranges: the hidden dimension in the LSTM is explored in {16, 32, 64}; $k$ is selected from {2, 3, 4, 5}; $a$ is tuned within {0.1, 0.2, 0.3, 0.4}; and the hypergraph depth varies among {1, 2, 3, 4}. Early stopping is utilized to avoid overfitting, with the patience set to 10. Additionally, $H^2$CAN is evaluated over five iterations with different random number seeds, and the mean results are taken.

## Baselines

To verify the effectiveness of $H^2$CAN, we conduct a comparison with the following methods.

**TFN** [4]. The model performs fusion by applying an outer product on tensors from different modalities. **LMF** [5]. To improve computational efficiency, the model decomposes high-order tensors into low-rank structures for multimodal fusion. **MFM** [52]. MFM leverages generative and discriminative representations to enhance feature learning across different modalities. **MulT** [8]. MulT utilizes a pairwise cross-modality attention mechanism to model inter-modality relationships. **MISA** [22]. The model captures both modal-specific and modal-general information, integrating orthogonal loss and reconstruction loss to optimize performance. **MAG-BERT** [53]. The model employs a Multimodal Adaptation Gate to enrich text representations with aligned visual and auditory features. **Self-MM** [18]. The model designs an unimodal label generation method for retaining modality-specific information. **BBFN** [54]. The model simultaneously performs modal fusion and information separation, and it incorporates a complementary mechanism to mitigate information imbalances between modalities. **MIB** [55] To filter noise from unimodal data, the model incorporates regularization to learn the representation effectively. **MUTA-Net** [56]. The model enhances the discrimination of representations by performing modal fusion at various levels and by designing a loss function that incorporates intra-class distance. **TeTFN** [9]. Based on the MulT architecture, TeTFN fuses information from nonverbal modalities from a text-based viewpoint. **HCIL** [41]. The model devises a hybrid cross-modality learning method to fully utilize multimodal sentiment information. **PEST** [24]. PEST utilizes a dynamic propagation method to capture inter-modality interactions of translated features. **VLP2MSA** [50]. VLP2MSA employs a fusion method that balances visual and text modalities to build multimodal representations.

## Overall performance evaluation and results

Tables 2 and 3 present the comparative results on benchmark datasets between our $H^2$CAN and baseline models. Additionally, we evaluate the impact of different encoders for non-text modalities by replacing LSTM with Transformer. The results indicate that LSTM and Transformer achieve comparable performance on the MSA task, potentially because the shallow features extracted by LSTM capture richer information [57]. On the MOSI, MOSEI, and CH-SIMS datasets, $H^2$CAN improves classification accuracy by approximately 1%, 1.2%, and 1.8%, respectively.

Specifically, $H^2$CAN exhibits significant performance advantages over traditional tensor fusion-based models, namely TFN and LMF. $H^2$CAN also shows improved results compared to models that focus on cross-modality attention fusion strategies, such as MulT, TeTFN, and VLP2MSA. On the MOSI dataset, the Acc-2 (negative/positive) and F1 (negative/positive) metrics improved from 86.28% and 86.28% for VLP2MSA to 87.20% and 87.17%, respectively, reflecting the advantage of our heterogeneous hypergraph-based fusion strategy. In addition, while VLP2MSA and TeTFN utilize attentional mechanisms during the modal fusion phase, our model further incorporates a CIT to enhance the attention module. Experimental results on the MOSI and MOSEI datasets illustrate the effectiveness of our Counterfactual

**Table 2** Comparison of $H^2$CAN and baseline model results on the MOSI and MOSEI datasets

| Model | CMU-MOSI | | | | CMU-MOSEI | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE↓ | Corr↑ | Acc-2↑ | F1-score↑ | MAE↓ | Corr↑ | Acc-2↑ | F1-score↑ |
| TFN[a](2017) | 0.901 | 0.698 | –/80.80 | –/80.70 | 0.593 | 0.700 | –/82.50 | –/82.10 |
| LMF[a](2018) | 0.917 | 0.695 | –/82.50 | –/82.40 | 0.623 | 0.677 | –/82.00 | –/82.10 |
| MFM[a](2018) | 0.877 | 0.706 | –/81.70 | –/81.60 | 0.568 | 0.717 | –/84.40 | –/84.30 |
| MulT[a](2019) | 0.861 | 0.711 | 81.50/84.10 | 80.60/83.90 | 0.580 | 0.703 | –/82.50 | –/82.30 |
| MISA[a](2020) | 0.783 | 0.761 | 81.80/83.40 | 81.70/83.60 | 0.555 | 0.756 | 83.60/85.50 | 83.80/85.30 |
| MAG-BERT[a](2020) | 0.727 | 0.781 | 82.37/84.43 | 82.50/84.61 | 0.543 | 0.755 | 82.51/84.82 | 82.77/84.71 |
| Self-MM[a](2021) | 0.713 | 0.798 | 84.00/85.98 | 84.42/85.95 | 0.530 | 0.765 | 82.81/85.17 | 82.81/85.30 |
| BBFN*(2022) | 0.776 | 0.755 | –/84.30 | –/84.30 | 0.561 | 0.731 | –/83.10 | –/83.20 |
| MIB*(2022) | 0.722 | 0.782 | –/85.30 | –/85.30 | 0.588 | 0.761 | –/85.40 | –/85.40 |
| MUTA-Net*(2023) | 0.730 | 0.793 | 83.10/85.00 | 83.00/85.00 | 0.544 | 0.760 | 82.40/85.00 | 82.70/84.90 |
| TeTFN*(2023) | 0.717 | 0.800 | 84.00/86.10 | 83.83/86.07 | 0.551 | 0.748 | 84.25/85.18 | 84.18/85.27 |
| HCIL*(2024) | 0.703 | 0.810 | 84.25/86.07 | 84.18/86.01 | 0.532 | 0.768 | 82.56/85.97 | 82.68/85.29 |
| PEST*(2024) | 0.723 | 0.796 | –/86.10 | –/86.10 | 0.542 | 0.764 | –/85.10 | –/85.30 |
| VLP2MSA*(2024) | **0.696** | **0.813** | 84.55/86.28 | 84.48/86.28 | 0.535 | 0.770 | 83.90/85.97 | 83.82/85.89 |
| $H^2$CAN + *trans* | 0.700 | 0.808 | **85.51**/86.95 | **85.43**/86.88 | **0.523** | 0.768 | 84.86/85.74 | 84.69/85.56 |
| $H^2$CAN (ours) | 0.704 ± | 0.805 ± | 85.42 ±/**87.20** ± | 85.35 ±/**87.17** ± | 0.530 ± | **0.770** ± | **85.08** ±/**86.45** ± | **85.08** ±/**86.27** ± |
| | 0.005 | 0.004 | 0.29/0.43 | 0.25/0.40 | 0.006 | 0.003 | 0.25/0.35 | 0.20/0.39 |
| p-value(t-test) | – | – | 0.0051/0.0035 | 0.0047/0.0043 | – | – | 0.0033/0.0039 | 0.0029/0.0038 |

For the Acc-2 and F1-score metrics, results for the 'negative/non-negative' and 'negative/positive' are split by '/'. + *trans* denotes replacing LSTM with Transformer in the model. Bolded results indicate the best performance. Values in the second row of '$H^2$CAN(ours)' represent the standard deviations of various metrics for the model across different random seeds. The 'p-value' reflects the statistical significance of comparisons with the publicly available method Self-MM

[a]Indicates that the results for the model are provided by [50]

*Indicates that the results for the model are sourced from its original paper

‾Indicates p-value $< 0.05$

Intervention Task. Compared to models like MFM, MISA, and Self-MM, which aim to capture both modality-specific and modality-independent information, our method achieves a 0.15 improvement in Corr and a 0.09 reduction in MAE. This highlights the superiority of the multimodal representations achieved by $H^2$CAN. Besides, we calculate the standard deviations and conduct paired t-tests to evaluate the improvements achieved by $H^2$CAN. The results show that the standard deviations fall within a narrow range, and the p-value in the t-tests is below 0.05, indicating that $H^2$CAN's performance is statistically significant compared to the baselines. These findings further confirm the effectiveness and superiority of our method.

Due to the complexity of Chinese semantics, on the CH-SIMS dataset, the models' performance is weaker compared to that achieved on the other two English datasets. Our model demonstrates a significant improvement on this dataset, with the F1-score increasing from 79.3 to 81.4%, validating the effectiveness of $H^2$CAN across different languages.

Perhaps our proposed $H^2$CAN model does not display a great advantage in Acc-2 (negative/positive) and F1 (negative/positive) metrics over other models on the MOSEI dataset. This can be attributed to the overrepresentation of samples with neutral sentiment in the dataset, which often lack diversity and limit the model's ability to fit other samples. This is confirmed by $H^2$CAN's performance, as it achieves the best results in both the Acc-2 (negative/non-negative) and F1 (negative/non-negative) metrics.

**Table 3** Results on the CH-SIMS dataset

| Models | MAE↓ | Corr↑ | Acc-2↑ | F1-score↑ |
|---|---|---|---|---|
| TFN[a] | 0.432 | 0.591 | 78.3 | 78.6 |
| LMF[a] | 0.441 | 0.576 | 77.7 | 77.8 |
| MulT[a] | 0.453 | 0.561 | 78.5 | 79.6 |
| VLP2MSA[a] | 0.412 | 0.604 | 79.4 | 79.3 |
| $H^2$CAN(ours) | **0.409** | **0.607** | **81.2** | **81.4** |

[a]Indicates that results are sourced from [50]

Bolded results indicate the best performance

## Effect of hyperparameters *k* and *α*

Considering the absence of explicit collaboration between the hyperparameters $k$ and $\alpha$, we restrict the values of $k$ and

**Table 4** Results for hyperparameter analysis of $k$ and $\alpha$

| Description | CMU-MOSI | | | | CMU-MOSEI | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-2 | F1-score | MAE | Corr | Acc-2 | F1-score |
| $k = 2$ | **0.695** | 0.805 | 84.39/86.58 | 84.21/86.47 | **0.530** | 0.770 | **85.08/86.45** | **85.08/86.27** |
| $k = 3$ | 0.704 | **0.805** | **85.42/87.20** | **85.35/87.17** | 0.532 | 0.765 | 84.46/85.50 | 84.52/85.29 |
| $k = 4$ | 0.713 | 0.797 | 84.26/86.59 | 84.17/86.45 | 0.534 | 0.758 | 83.75/85.50 | 83.85/85.27 |
| $k = 5$ | 0.720 | 0.791 | 84.26/86.13 | 84.03/86.12 | 0.540 | 0.761 | 83.43/85.44 | 83.72/85.18 |
| $\alpha = 0.1$ | 0.710 | 0.798 | 82.94/85.82 | 82.64/85.65 | **0.530** | **0.770** | **85.08/86.45** | **85.08/86.27** |
| $\alpha = 0.2$ | 0.712 | 0.800 | 83.67/86.28 | 83.88/86.07 | 0.537 | 0.762 | 83.64/85.11 | 83.85/84.99 |
| $\alpha = 0.3$ | **0.704** | **0.805** | **85.42/87.20** | **85.35/87.17** | 0.541 | 0.759 | 83.75/84.78 | 83.94/84.46 |
| $\alpha = 0.4$ | 0.724 | 0.794 | 83.82/85.98 | 83.69/85.95 | 0.535 | 0.760 | 84.03/85.17 | 84.16/85.02 |

Bolded results indicate the best performance

$\alpha$ independently to explore their respective effects on the model. Table 4 presented the result of these experiments.

The hyperparameter $k$ adjusts the range of audiovisual nodes included in each hyperedge, and a smaller value of $k$ indicates a broader range of interaction between nodes. As the value of $k$ increases, the performance of $H^2$CAN on the MOSI dataset initially rises, peaking at a $k$ value of 3, and then decreases. Conversely, on the MOSEI dataset, it reaches an optimum at a $k$ value of 2 and subsequently continues to decline. This demonstrates that with sufficiently large datasets, our proposed fusion strategy benefits from a broader range of node interactions.

The hyperparameter $\alpha$ is used to adjust the weighting ratio of attention scores corresponding to the biased modality in the counterfactual fusion branch. By increasing the value of $\alpha$, we magnify the effects of bias information in the counterfactual fusion branch. On the MOSI dataset, the performance of $H^2$CAN initially increases and then decreases with rising $\alpha$ values, reaching its optimum at an $\alpha$ value of 0.3. Conversely, on the MOSEI dataset, as $\alpha$ increases, the performance of $H^2$CAN continues to decline.

## Ablation studies

Comprehensive ablation studies are conducted to verify the effectiveness of different modules in $H^2$CAN. The details of these ablation studies are provided as follows.

**w/ KNN**. We use the KNN algorithm to construct a dynamic heterogeneous hypergraph.

**w/ split**. We split the original hyperedge, which includes three modalities, into two separate hyperedges: one for text and audio and another for text and visual.

**w/ HGAT**. We replace heterogeneous hypergraphs with heterogeneous graphs, utilizing the heterogeneous graph attention network (HGAT) to capture inter-modality interactions.

**w/o TF**. Neglecting the interactions of text with other modalities.

**w/o AF**. Neglecting the interactions of audio with other modalities.

**w/o VF**. Neglecting the interactions of visual with other modalities.

**w/o CIT**. We remove the Counterfactual Intervention Task during the training phase of our model.

**w/o ULGM**. We remove the ULGM and randomly select modality for additional weighting.

**w/o AW**. We generate counterfactual attention scores using predefined distributions without employing additional weighting for modality attention scores.

### Effect of heterogeneous hypergraph depth

Figure 3 illustrates the impact of different heterogeneous hypergraph depths on the model's performance across two datasets. The overall performance of $H^2$CAN initially increases with the depth of the hypergraph, then decreases, ultimately reaching an optimum at a depth of three across both datasets. This is because, with a small number of HHGAT layers, the model fails to adequately capture the complex interaction information between modalities, while an excessively large number of layers can lead to overfitting.

### Effect of heterogeneous hypergraph construct method

Table 5 presents the effects of different heterogeneous hypergraph construction methods. Firstly, we utilize the KNN-based construction method, which is commonly used to construct heterogeneous hypergraphs. From the comparison results, our designed heterogeneous hypergraph construction method proves to be superior to the KNN-based construction method. This suggests that the KNN algorithm is not well-suited for heterogeneous multimodal data, whereas the construction method we devised, which incorporates the

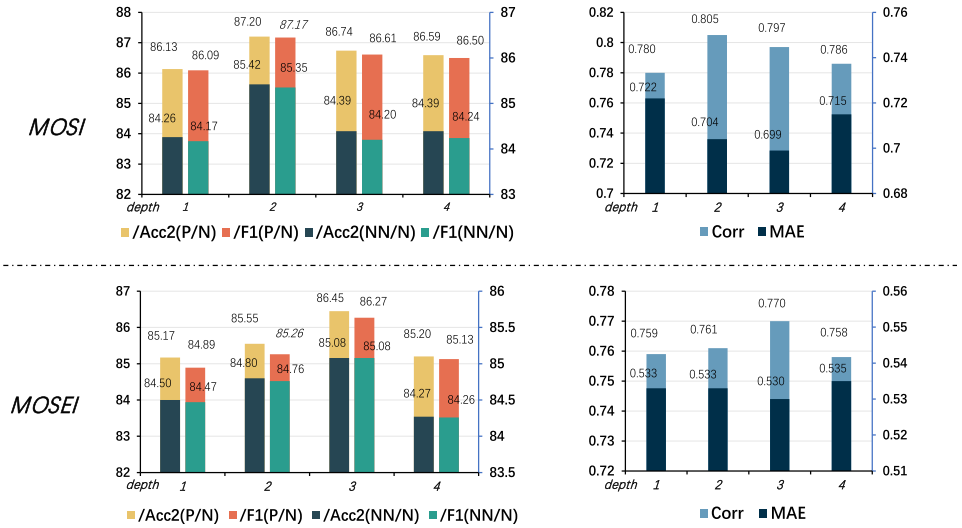


**Fig. 3** Results under different depths of heterogeneous hypergraphs. Notably, each bar graph has two vertical axes

**Table 5** Results for different heterogeneous hypergraph construct methods on the MOSI and MOSEI datasets

| | CMU-MOSI | | | |
|---|---|---|---|---|
| | MAE | Corr | Acc-2 | F1-score |
| $H^2$CAN | **0.704** | **0.805** | **85.42/87.20** | **85.35/87.17** |
| w/ KNN | 0.721 | 0.797 | 83.82/85.06 | 83.79/85.00 |
| w/ split | 0.718 | 0.792 | 84.11/85.21 | 84.03/85.18 |
| | CMU-MOSEI | | | |
| | MAE | Corr | Acc-2 | F1-score |
| $H^2$CAN | **0.530** | **0.770** | **85.08/86.45** | **85.08/86.27** |
| w/ KNN | 0.539 | 0.757 | 83.54/84.23 | 83.72/83.88 |
| w/ split | 0.537 | 0.754 | 83.37/84.42 | 84.54/84.27 |

Bolded results indicate the best performance

characteristics of sentiment expression, is more suitable for the MSA task. Moreover, compared to the KNN-based construction method, our construction method is static and requires fewer computational resources. We then split the original three-modal hyperedge into two bimodal hyperedges to facilitate further exploration. From Table 5, we can observe that the original tri-modality hypergraph construction method performs better. This is because the inter-modal interactions modeled by the tri-modality hyperedge are more comprehensive than those modeled by the two bi-modality hyperedges.

### Effect of the fusion strategy

Figure 4 presents the ablation experiment results of our heterogeneous hypergraph attention-based fusion strategy. From Fig. 4a and c, the positive impact of modeling inter-modality beyond-pairwise interactions through heterogeneous hypergraphs is evident. Compared with traditional graph attention-based fusion strategies, our designed heterogeneous hypergraph attention-based fusion strategy yields an improvement of more than 1% in both the Acc-2 and F1 metrics. Furthermore, Fig. 4b and d illustrate that the removal of interactions between any modality results in a performance decrease. Therefore, we can conclude that our designed fusion strategy effectively models the beyond-pairwise interactions among multiple nodes within multiple modalities through heterogeneous hypergraph attention networks.

### Effect of counterfactual intervention task

In Fig. 5, we present experiments that involve removing the CIT and modifying its implementation details to assess their impact. Firstly, we find that incorporating the CIT provides the model with a performance improvement of 0.6–0.9% in both the F1 and Acc-2 metrics. This suggests that the CIT effectively enhances model performance by supervising attention learning. Secondly, we observe that either randomly selecting a modality for additional weighting or canceling the additional weighting strategy for modality can adversely affect $H^2$CAN. This indicates that appropriately weighting the attentional scores corresponding to the modality helps to mitigate the effects of bias information.

## Feature visualization

As displayed in Fig. 6, we perform a t-SNE visualization on the MOSI dataset to explore the model's feature learning performance under various fusion strategies. Comparing

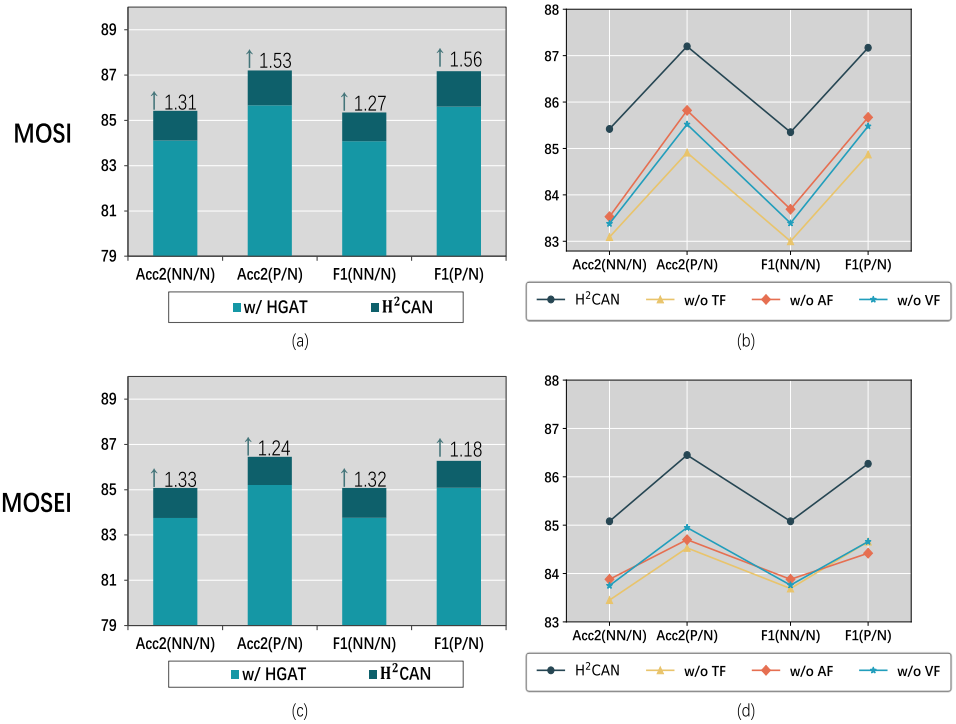


**Fig. 4** Results for different fusion strategy

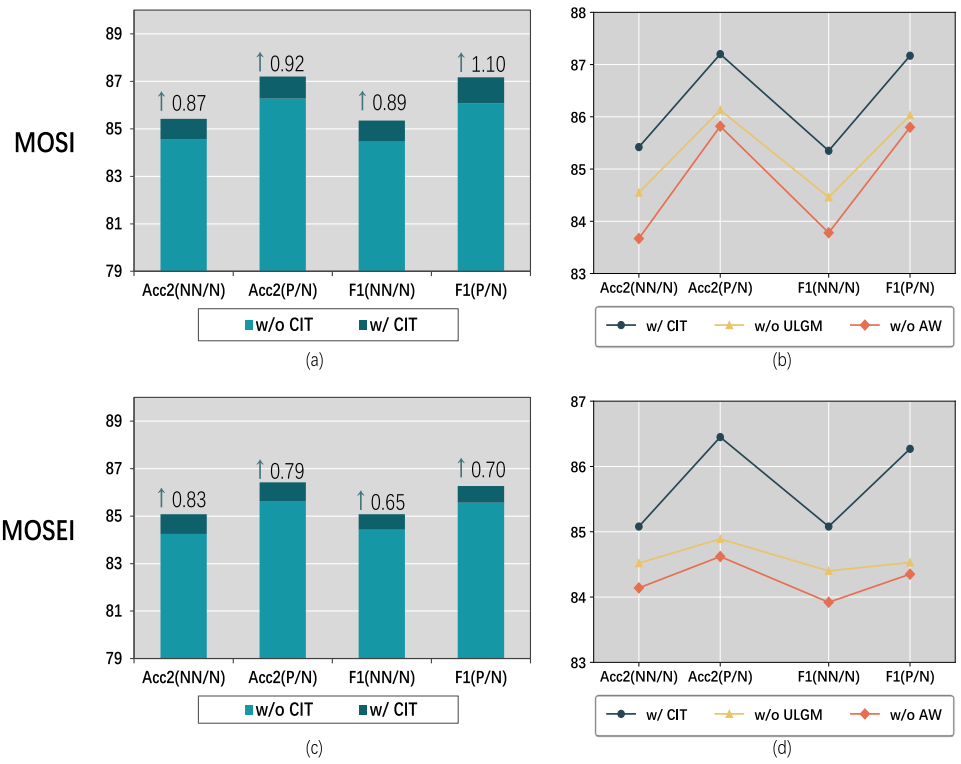


**Fig. 5** Results for counterfactual intervention task

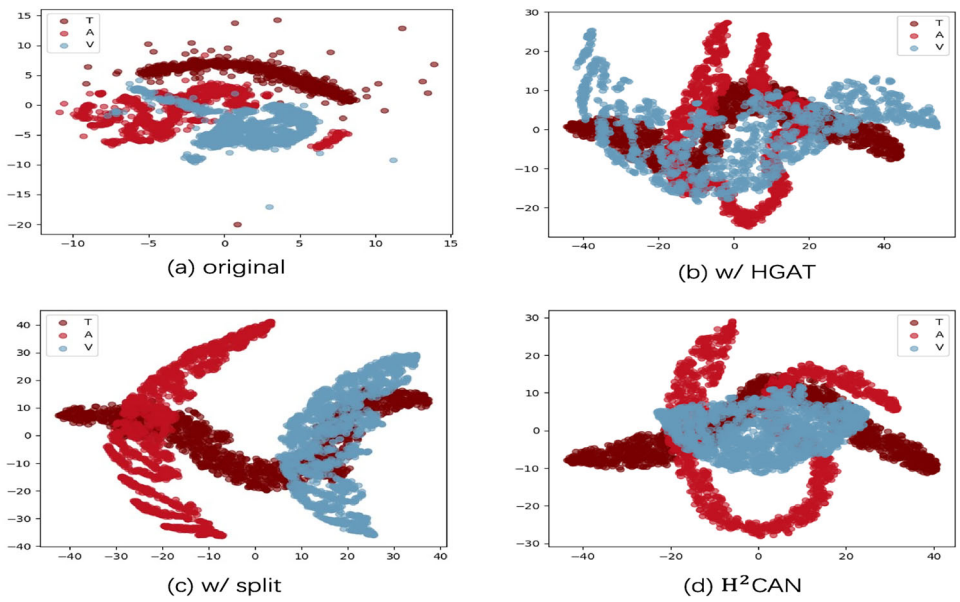


**Fig. 6** The T-SNE algorithm is used to generate the visualization of unimodal features in the MOSI dataset. The text, audio, and visual modalities are denoted by 'T', 'A', and 'V', respectively. **a** Shows the features before counterfactual modal fusion. **b** Shows the features obtained by modeling inter-modality interactions using HGAT. **c** Shows the features obtained by modeling inter-modality interactions with HHGAT after splitting the tri-modality hyperedge into text-audio and text-visual hyperedges. **d** Shows the features obtained by modeling inter-modality interactions using HHGAT

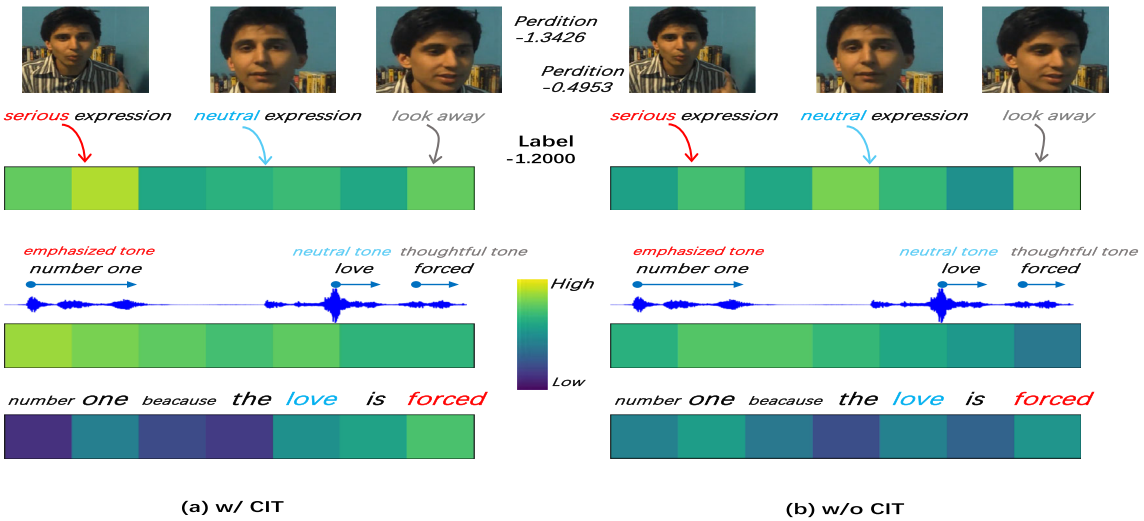


**Fig. 7** Heat map visualization of attention scores during the first stage of message passing in the last layer of HHGAT. **a** and **b** display the visualization results with and without the Counterfactual Intervention Task, respectively. 'Label' refers to the multimodal sentiment intensity label of the sample. The color of the heat map from dark to light indicates low to high attention scores

the visualization results before and after modeling inter-modality interactions, we find that feature clustering can be effectively achieved through this method. We can observe from Fig. 6b and c that the features obtained from our fusion strategy are more concentrated and have more distinct boundaries. In addition, we believe that fusion using tri-modality hyperedges can capture richer inter-modality interaction information. This is evidenced by the larger spacing between the audio and visual feature clusters in Fig. 6c compared to Fig. 6d, which utilize bi-modality hyperedges and tri-modality hyperedges, respectively. These comparative results illustrate that our designed fusion strategy not only effectively combines information from different modalities but also preserves modality-specific information.

## Case study

To more deeply explore the effect of CIT, we selected a sample from the validation set of the MOSI dataset and visualized its attention scores, as shown in Fig. 7. For convenience, we apply average pooling to the attention scores obtained from different hyperedges.

From the attention scores in Fig. 7a and b, we can intuitively observe that the attention of the model with CIT captures more accurate inter-modal interactions. Specifically,

**Table 6** The parameter count and FLOPs of both $H^2$CAN and other state-of-the-art models

| Models | Params(M) | FLOPs(M) | MOSI-Acc-2 |
|---|---|---|---|
| MAG-BERT | 86.947 | 17.256 | 82.48/84.02 |
| TeTFN | 87.136 | 34.531 | 84.05/86.10 |
| $H^2$CAN | 86.697 | 17.120 | 85.42/87.20 |

for the textual modality, the difference in attention scores between the words 'love' and 'forced' becomes subtle after the removal of the Counterfactual Intervention Task. This suggests that the model may be misled by bias information from the word 'love'. Similar results are observed in the audio and visual modalities. For instance, in Fig. 7b, the 'neutral expression' is assigned a relatively high weight, and the 'emphasized tone' is given a relatively low weight. The prediction of the model with CIT is much closer to the label, indicating that CIT guides the model's attention learning to mitigate the effects of bias information and improve the model's performance.

### Model complexity

To evaluate the complexity of our model, we conduct a comparison of parameter counts and FLOPs between $H^2$CAN and other state-of-the-art approaches. As shown in Table 6, $H^2$CAN achieves a slight advantage over these models in both parameters and FLOPs. Additionally, on the CMU-MOSI dataset, $H^2$CAN achieves an accuracy increase of 1.4% compared to these models, illustrating that our method effectively balances performance with complexity.

## Conclusion

In this paper, we proposed a Heterogeneous Hypergraph Attention Network with Counterfactual Learning to address the limitations in cross-modality attention methods within existing MSA models. Based on the sentiment expression characteristics of different modalities, we designed a static hypergraph construction method and employed hypergraph convolution using the attention mechanism to model the beyond-pairwise interactions among modalities. Furthermore, we designed a Counterfactual Intervention Task based on causal reasoning to guide attention learning by maximizing the difference between actual and counterfactual predictions, thereby mitigating the interference of bias information. Related experiments were conducted on the benchmark datasets to demonstrate the significant improvement of our model compared to recent state-of-the-art models. Additionally, we performed visualization and case studies to intuitively explore the effects of hypergraph-based fusion strategies and the Counterfactual Intervention Task.

As our model fails to draw on the latest research on modality translation, it suffers from modality heterogeneity when capturing inter-modality interactions. In subsequent work, we will attempt to develop new methods to reduce the impact of modality heterogeneity, thereby improving the performance of multimodal models.

**Acknowledgements** This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2022C03106) and the National Natural Science Foundation of China (No. 62037001, No. 62337001).

**Author Contributions** Changqin Huang: conceptualization, visualization, writing—review and editing. Zhenheng Lin: data curation, writing—original draft, visualization. Qionghao Huang: supervision, funding acquisition, writing—review and editing. Xiaodi Huang: writing—review and editing. Fan Jiang: writing—review and editing, Jili Chen: writing—review and editing.

**Data Availability** Data are available on request from the authors.

## Declarations

**Conflict of interest** All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. On behalf of all authors, the corresponding author states that there is no conflict of interest.



## References

1. Li Z, Zou Z (2024) Punctuation and lexicon aid representation: a hybrid model for short text sentiment analysis on social media platform. J King Saud Univ Comput Inf Sci 36(3):102010
2. Chandrasekaran G, Nguyen TN, Hemanth DJ (2021) Multimodal sentimental analysis for social media applications: a comprehensive review. Wiley Interdiscip Rev: Data Min Knowl Discov 11(5):e1415
3. Huang J, Zhou J, Tang Z, Lin J, Chen CYC (2024) TMBL: transformer-based multimodal binding learning model for multimodal sentiment analysis. Knowl Based Syst 285:111346
4. E, Morency LP (2017) Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on

Empirical Methods in Natural Language Processing. Pittsburgh: Association for Computational Linguistics, pp 1103–1114
5. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, 853 Morency LP (2018) 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). Pittsburgh: Association for Computational Linguistics, pp 2247–2256
6. Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency LP (2018) Multi-attention recurrent network for human communication comprehension. In: In Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, pp 5642–5649
7. Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A (2023) Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf Fusion 91:424–444
8. Tsai YHH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R (2019) Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for computational linguistics. Meeting, vol 2019. New York: NIH Public Access, p 6558
9. Wang D, Guo X, Tian Y, Liu J, He L, Luo X (2023) TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognit 136:109259
10. Xiao L, Wu X, Wu W, Yang J, He L (2022) Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). New York: IEEE, pp 4578–4582
11. Huang J, Pu Y, Zhou D, Cao J, Gu J, Zhao Z et al (2024) Dynamic hypergraph convolutional network for multimodal sentiment analysis. Neurocomputing 565:126992
12. Keltner D, Sauter D, Tracy J, Cowen A (2019) Emotional expression: advances in basic emotion theory. J Nonverbal Behav 43:133–160
13. Zhang F, Li XC, Lim CP, Hua Q, Dong CR, Zhai JH (2022) Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. Inf Fusion 88:296–304
14. Rao Y, Chen G, Lu J, Zhou J (2021) Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, pp 1025–1034
15. Zhu Y, Chen Z, Wu F (2019) Multimodal deep denoise framework for affective video content analysis. In: Proceedings of the 27th ACM International Conference on Multimedia. New York: Association for Computing Machinery, pp 130–138
16. Kotsia I, Pitas I (2005) Real time facial expression recognition from image sequences using support vector machines. In: IEEE international conference on image processing 2005, vol 2. New York: IEEE, pp II–966
17. Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency LP (2019) Words can shift: dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, Palo Alto: AAAI Press, pp 7216–7223
18. Yu W, Xu H, Yuan Z, Wu J (2021) 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, Palo Alto: AAAI Press, pp 10790–10797
19. Zhang X, Wang Z, Cao G, Ho SB (2024) Joint weakly supervised image emotion analysis based on interclass discrimination and interclass correlation. IEEE Intell Syst 39(5):82–89
20. Zhang H, Liu Y, Xiong Z, Wu Z, Xu D (2024) Visual sentiment analysis with semantic correlation enhancement. Complex Intell Syst 10(2):2869–2881
21. Yang J, Yu Y, Niu D, Guo W, Xu Y (2023) ConFEDE: contrastive feature decomposition for multimodal sentiment analysis. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, vol 1 (long papers), Pittsburgh: Association for Computational Linguistics, pp 7617–7630
22. Hazarika D, Zimmermann R, Poria S (2020) Misa: modalityinvariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia, New York: Association for Computing Machinery, pp 1122–1131
23. Huang C, Zhang J, Wu X, Wang Y, Li M, Huang X (2023) TeFNA: text-centered fusion network with crossmodal attention for multimodal sentiment analysis. Knowl Based Syst 269:110502
24. Gan C, Tang Y, Fu X, Zhu Q, Jain DK, García S (2024) Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation. Knowledge-Based Systems, 299: 111982
25. Wang L, Peng J, Zheng C, Zhao T et al (2024) A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. Inf Process Manag 61(3):103675
26. Yang J, Wang Y, Yi R, Zhu Y, Rehman A, Zadeh A et al (2020) MTAG: modal-temporal attention graph for unaligned human multimodal language sequences. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Pittsburgh: Association for Computational Linguistics, pp 1009–1021
27. Feng Y, You H, Zhang Z, Ji R, Gao Y (2019) Hypergraph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, Palo Alto: AAAI Press, pp 3558–3565
28. Kim ES, Kang WY, On KW, Heo YJ, Zhang BT (2020) Hypergraph attention networks for multimodal learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New York: IEEE, pp 14581–14590
29. Li Q, Ji C, Guo S, Zhao Y, Mao Q, Wang S et al (2024) Variational multi-modal hypergraph attention network for multi-modal relation extraction. In: Proceedings of the ACM International Conference on Multimedia 2024. New York: Association for Computing Machinery, pp 1–11
30. Khan B, Wu J, Yang J, Ma X (2023) Heterogeneous hypergraph neural network for social recommendation using attention network. In: ACM Transactions on Recommender Systems. New York: Association for Computing Machinery, pp 1–23
31. Das R, Singh TD (2023) Multimodal sentiment analysis: a survey of methods, trends, and challenges. ACM Comput Surv 55(13s):1–38
32. Li S, Guo D, Liu K, Hong R, Xue F (2023) Multimodal counterfactual learning network for multimedia-based recommendation. In: Proceedings of the 46th international ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Association for Computing Machinery, pp 1539–1548
33. Zhang X, Wang Z, Wang H, Xiang J, Wu C, Cao G (2024) CausVSR: causality inspired visual sentiment recognition. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann Publishers Inc., pp 3196–3204
34. Sun T,WangW, Jing L, Cui Y, Song X, Nie L (2022) Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In: Proceedings of the 30th ACM International Conference on Multimedia. New York: Association for Computing Machinery, pp 15–23
35. Huang C, Chen J, Huang Q, Wang S, Tu Y, Huang X (2025) AtCAF: attention-based causality-aware fusion network for multimodal sentiment analysis. Inf Fusion 114:102725
36. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training 960 of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North Ameri-

can Chapter of the Association for Computational Linguistics: Human Language Technologies. Pittsburgh: Association for Computational Linguistics, pages 4171–4186
37. Zhu Q, Yeh MC, Cheng KT, Avidan S (2006) Fast human detection using a cascade of histograms of oriented gradients. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol 2. New York: IEEE, pp 1491–1498
38. Degottex G, Kane J, Drugman T, Raitio T, Scherer S (2014) COVAREP—a collaborative voice analysis repository for speech technologies. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). New York: IEEE, pp 960–964
39. Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint. arXiv:1508.01991
40. Vaswani A (2017) Attention is all you need. In: Advances in Neural Information Processing Systems. San Mateo: Morgan Kaufmann Pubs, pp 1–12
41. Fu Y, Zhang Z, Yang R, Yao C (2024) Hybrid cross-modal interaction learning for multimodal sentiment analysis. Neurocomputing 571:127201
42. You Y, Chen T, Sui Y, Chen T, Wang Z, Shen Y (2020) Graph contrastive learning with augmentations. Adv Neural Inf Process Syst 33:5812–5823
43. Shipley B (2016) Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R. Cambridge University Press, Cambridge
44. Pearl J (2022) Direct and indirect effects. In: Probabilistic and causal inference: the works of Judea Pearl. New York: Association for Computing Machinery, pp 373–392
45. Chang CH, Adam GA, Goldenberg A (2021) Towards robust classification model by counterfactual and invariant data generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New York: IEEE, pp 15212–15221
46. VanderWeele T (2015) Explanation in causal inference: methods for mediation and interaction. Oxford University Press, Oxford
47. Zadeh A, Zellers R, Pincus E, Morency LP (2016) Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell Syst 31(6):82–88
48. Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP (2018) 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, vol 1 (long papers). Pittsburgh: Association for Computational Linguistics, pp 2236–2246
49. Yu W, Xu H, Meng F, Zhu Y, Ma Y, Wu J et al (2020) CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Pittsburgh: Association for Computational Linguistics, pp 3718–3727
50. Yi G, Fan C, Zhu K, Lv Z, Liang S, Wen Z et al (2024) VLP2MSA: expanding vision-language pre-training to multimodal sentiment analysis. Knowl Based Syst 283:111136
51. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint. arXiv:1412.6980
52. Tsai YHH, Liang PP, Zadeh A, Morency LP, Salakhutdinov R (2018) Learning factorized multimodal representations. arXiv preprint. arXiv:1806.06176
53. Rahman W, Hasan MK, Lee S, Zadeh A, Mao C, Morency LP et al (2020) Integrating multimodal information in large pretrained transformers. In: Proceedings of the conference. Association for computational linguistics. Meeting, vol 2020. New York: NIH Public Access, p 2359
54. Han W, Chen H, Gelbukh A, Zadeh A, Morency L, Poria S (2021) Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In: Proceedings of the 2021 International Conference on Multimodal Interaction. New York: Association for Computing Machinery, pp 6–15
55. Mai S, Zeng Y, Hu H (2022) Multimodal information bottleneck: learning minimal sufficient unimodal and multimodal representations. IEEE Trans Multimed 25:4121–4134
56. Tang Z, Xiao Q, Zhou X, Li Y, Chen C, Li K (2023) Learning discriminative multi-relation representations for multimodal sentiment analysis. Inf Sci 641:119125
57. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, pp 770–778