Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis

Teng Sun
Shandong University
stbestforever@gmail.com

Wenjie Wang*
National University of Singapore
wenjiewang96@gmail.com

Liqiang Jing
Shandong University
jingliqaing6@gmail.com

Yiran Cui
Shandong University
conyrol120@gmail.com

Xuemeng Song
Shandong University
sxmustc@gmail.com

Liqiang Nie*
Harbin Institute of Technology (Shenzhen)
nieliqiang@gmail.com

ABSTRACT

Existing studies on multimodal sentiment analysis heavily rely on textual modality and unavoidably induce the spurious correlations between textual words and sentiment labels. This greatly hinders the model generalization ability. To address this problem, we define the task of out-of-distribution (OOD) multimodal sentiment analysis. This task aims to estimate and mitigate the bad effect of textual modality for strong OOD generalization. To this end, we embrace causal inference, which inspects the causal relationships via a causal graph. From the graph, we find that the spurious correlations are attributed to the direct effect of textual modality on the model prediction while the indirect one is more reliable by considering multimodal semantics. Inspired by this, we devise a model-agnostic counterfactual framework for multimodal sentiment analysis, which captures the direct effect of textual modality via an extra text model and estimates the indirect one by a multimodal model. During the inference, we first estimate the direct effect by the counterfactual inference, and then subtract it from the total effect of all modalities to obtain the indirect effect for reliable prediction. Extensive experiments show the superior effectiveness and generalization ability of our proposed framework.

CCS CONCEPTS

• **Information systems → Multimedia information systems**.

KEYWORDS

Multimodal Sentiment Analysis, Counterfactual Reasoning, Out-of-distribution Generalization

ACM Reference Format:
Teng Sun, Wenjie Wang, Liqiang Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. 2022. Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisbon, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548211

*Wenjie Wang (wenjiewang96@gmail.com) and Liqiang Nie (nieliqiang@gmail.com) are corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '22, October 10–14, 2022, Lisbon, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548211

1 INTRODUCTION

Sentiment analysis, as a typical language understanding task, aims to automatically recognize the subjective opinions from users' reviews (*e.g.*, positive, neutral, or negative attitudes). Due to its wide applications in social media [28] and recommender systems [45], sentiment analysis has attracted extensive attention from both academia and industry. Early work mainly relies on the textual information for sentiment analysis [2]. However, with the boom of mobile social media, more users tend to present their opinions in more diverse modalities, like images, audios, and videos. For example, facial expressions and the tone of voice in videos usually express emotional tendencies, which alleviate the ambiguity of textual modality. Inspired by this, the research focus has been recently shifted from the textual sentiment analysis to the multimodal one [42].

Existing research on Multimodal Sentiment Analysis (MSA) mainly aims to mitigate the modality gap and learn discriminative multimodal representations. According to the fusion manner, they generally fall into two categories: early-fusion and late-fusion methods. The former integrates multimodal features in the shallow layers of neural networks before reasoning. For example, Tsai *et al.* [29] fused the raw features of different modalities with cross-modal attentive mechanisms, and then recognized the sentiments from the fused embeddings. By contrast, the latter learns the intra-modal representation separately, and performs the cross-modal analysis independently. For example, Liu *et al.* [16] first extracted the unimodal representation from each audio, video and text by three sub-networks without cross-modal interactions, and then leveraged a low-rank model to fuse the sentiment analysis results yielded by each modality.

Despite their compelling progress, existing studies usually suffer from fitting the spurious correlations in textual modality [8]. As shown in Figure 1(b), some words in textual modality seem strongly correlated with the sentiment label, which however are likely to be unreliable. For example, in the training set, the word "movie" appears in negative samples more frequently than that in the positive ones, making the model more likely to classify the testing samples with "movie" as negative. Indeed, "movie" is not a reliable cue for identifying negative samples. Such correlations are known as spurious correlations [20] in causality, showing the bad effect of the textual modality. Deep neural networks are easy to adapt such

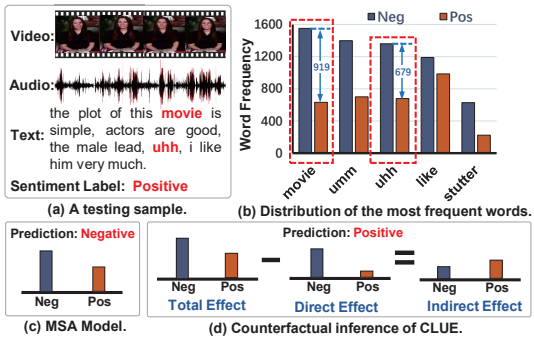


Figure 1: A testing sample, the distribution of the top-5 most frequent words in MOSEI [40], and the predictions of MSA model and CLUE for the testing sample.

correlations for prediction due to the short-cut bias [5], degrading the generalization ability in the out-of-distribution (OOD) scenarios where the correlations between textual words and labels differ from those in the training set. To measure the true generalization ability of the model, we propose a novel OOD MSA task to evaluate whether MSA models can reduce the bad effect of textual modality in the OOD scenarios.

To alleviate the bad effect, an intuitive idea is to collect an unbiased dataset without spurious correlations. However, it is infeasible because various correlations between extensive words and the sentiment labels commonly exist in the real world. Worse still, we cannot abandon the textual modality for sentiment analysis, considering its rich sentimental cues. In this light, our objective turns to eliminate the bad effect caused by spurious correlations while keeping the valuable semantics in textual modality. The keys to achieve this objective are threefold: 1) disentangling the good and bad effects of textual modality on the model prediction, 2) mitigating the bad effect for stronger OOD generalization, and 3) utilizing multimodal cues to alleviate the textual correlations.

To this end, we resort to causal inference [6]. In particular, we utilize a causal graph to inspect the causal relationships between multiple modalities and the model prediction. From the graph, we find that multimodal cues are the key to disentangle the good and bad effects. As described in Section 4.2, we discover two kinds of effects of textual modality on the model prediction: direct and indirect. The former implies a shortcut from textual words to the prediction, exhibiting the spurious correlations in textual modality (*i.e.,* the bad effect). As for the latter, it extracts reliable textual semantics by considering the multimodal cues. For example, the excited tone and the smiling face in the video well guide the model to focus on the word "great" and filter out the spurious correlations caused by "movie". As such, the indirect effect learned from the cross-modal alignment represents the good effect of texts. Traditional MSA models are blindly optimized, resulting in correlation fitting between multiple modalities and sentiment labels, whereby two effects are mixed and the optimization hence inevitably favors the spurious correlations in the shortcut [5].

To disentangle the two effects and mitigate the bad one, we devise a model-agnostic CounterfactuaL mUltimodel sEntiment (CLUE) framework. Specifically, CLUE calculates the indirect effect of textual modality by considering cross-modal modeling with a MSA model. Meanwhile, CLUE estimates the direct effect by adding an extra text model during training. In the inference stage, CLUE estimates the direct effect by counterfactual inference, which imagines *what the prediction would be if the model had only seen the effect of texts.* Thereafter, CLUE mitigates such a direct effect and utilizes the indirect one of multiple modalities to alleviate the influence of spurious correlations in texts. To evaluate the generalization ability, we construct the OOD testing set with significant correlation shifts. Besides, we instantiate our CLUE framework on several latest MSA models. The empirical results on two benchmarks show that CLUE is able to significantly strengthen the generalization ability of the MSA models on the OOD testing set while maintaining competitive performance on the biased one.

To sum up, our contributions are threefold.

- We define a novel OOD MSA task, which points out the spurious correlations in textual modality and highlights the necessity of strong OOD generalization abilities.
- We devise a model-agnostic CLUE framework. It strengthens the existing MSA models via capturing the causal relationships in the training set and mitigating the bad effect of textual modality by the counterfactual inference.
- We conduct extensive experiments on two benchmark datasets, and the results demonstrate the superior effectiveness and generalization ability of CLUE. As a byproduct, we have released the source code to facilitate the reproduction[1].

2 RELATED WORK

• **Multimodal Sentiment Analysis.** The goal of MSA is to regress or classify the overall sentiment of an utterance via acoustic, visual, and textual cues. Over the past decades, statistical knowledge and shallow machine learning algorithms have been frequently utilized in MSA [7, 11]. As shallow machine learning is limited by feature engineering and massive manual work for data annotation, deep learning becomes the mainstream technique for MSA [12, 33, 36, 43] in an end-to-end manner [38].

In a sense, the key to developing deep learning-based MSA models lies in two aspects: multimodal representation learning and cross-modal fusion. Studies on the former are in three variants: 1) Translation-based models translate one modality to another[17, 22, 23]. 2) Correlation-based models learn cross-modal correlations by using Canonical Correlation Analysis [1]. And 3) the shared subspace learning models map all the modalities simultaneously into a new shared subspace via specific algorithms, like adversarial learning [18]. Another line of MSA focuses on developing sophisticated fusion techniques. So far, the existing fusion strategies can be categorized into feature-level and decision-level fusion. The former focus on fusing the representations of different modalities with the help of methods such as the tensor fusion network [39]. Regarding the decision-level fusion, the features of different modalities are classified independently based on their individual features, and then all classification results are integrated into a final prediction [3].

Although existing studies have achieved great success, they ignore the spurious correlations between texts and the sentiment

[1]https://msa-clue.wixsite.com/clue.

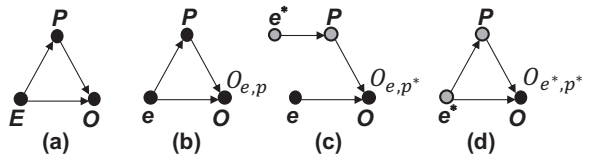


Figure 2: (a) The causal graph of the admission outcome, where the admission outcome of graduate (O) is directly affected by the experience (E) and publication (P). Counterfactual notations $O_{e,p}$, O_{e,p^*} and O_{e^*,p^*} are illustrated in (b), (c) and (d), respectively.

labels. As such, we proposed a new task, named OOD MSA, with a counterfactual framework to improve the OOD generalization ability of MSA models.

• **Causal Inference.** Causal inference has been widely-used across many tasks to improve the robustness of deep learning models, including visual question answering [13, 14, 19], visual commonsense reasoning [44], recommended system [32], and text classification [24]. Regarding the theory of causal inference, counterfactual reasoning enables models to achieve counterfactual imagination and estimate the causal effect. The methods of counterfactual reasoning can be broadly categorized into three strategies: 1) Some efforts estimated the path-specific causal effect by counterfactual inference such as natural direct effect and total indirect effect [19], and then mitigate some effects to achieve unbiased prediction [31]. 2) Some researchers trained a "poisonous" model regardless of the dataset biases and utilized *do* calculus to make unbiased predictions in the test phase [24]. 3) Existing studies also consider generating counterfactual samples for a factual sample and compare the counterfactual and factual samples to retrospect the original model prediction [4].

As for the MSA task, the model is at risk of capturing the bias from the textual modality and may make predictions based on the spurious correlations between texts and labels. Therefore, different from existing work, we attempted to apply counterfactual reasoning to MSA for mitigating the spurious correlations.

3 PRELIMINARY

In this section, we refer to this work [19] to detail the key concepts and symbols related to causal inference.

Casual graph is a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ which reveals the causal relationships among variables. Meanwhile, the nodes $\mathcal{N}$ denote the variables and the edges $\mathcal{E}$ represent the causal relationships among variables. An example of a causal graph about the admission outcome is illustrated in Figure 2(a). The casual graph mainly reflects the relationships among the experience (E), the publication (P), and the admission outcome (O): 1) O is directly affected by E and 2) indirectly influenced by E via the mediator P.

Counterfactual Inference. The counterfactual inference is to answer the counterfactual question based on factual observations, *i.e., "what the outcome would be if the treatment was different from the factual one?"* [20]. The treatment can be seen as assigning a specific value to the random variable. We utilize the lowercase letters to denote the specific values of random variables. Furthermore, we introduce structural equations [20] to formulate the causal relationships among causal variables for the convenience of subsequent understanding. For example, the values P and O under the conditions of $E = e$ and $P = p$ can be formulated as follows,

$$P_e = p = f_P(E = e), O_{e,p} = f_O(E = e, P = p), \tag{1}$$

where P_e denotes what the publication of someone would be if he/she has the experience e , and $O_{e,p}$ indicates what the admission outcome of someone would be if he/she has the experience e and publication p (see Figure 2(b)). $f_P(\cdot)$ and $f_O(\cdot)$ denote the structural equations for the variables P and O, respectively, which can be learned from the observational data [20].

Then, we can describe the counterfactual and factual situations by notations in the structural equations. For example, Figure 2(c) illustrates "what the admission outcome would be if someone only has the publication without the experience?", where O receives $E = e$ by $E \rightarrow O$, while P receives $E = e^*$ through $E \rightarrow P$. $E = e^*$ denotes no experience whatsoever, otherwise, $E = e$.

Causal effect. Causal effect refers to the difference in the outcome variable when the reference value of the variable change to a specific one [26], it is also named as Total Effect (TE). For example, TE of the experience $E = e$ (*i.e.,* the treatment variable) on the admission outcome O can be formulated as follows,

$$\begin{cases} \text{TE} = O_{e,p} - O_{e^*,p^*} = f_O(E = e, P = p) - f_O(E = e^*, P = p^*), \\ p^* = P_{e^*} = f_P(E = e^*), \end{cases} \tag{2}$$

where O_{e^*,p^*} denotes the reference situation, E and P are set as the reference values e^* (*i.e.,* no experience) and p^* (*i.e.,* the publication without experience), respectively. The corresponding illustration in the causal graph is presented in Figure 2(d). Specifically, O_{e^*,p^*} denotes what the admission outcome of someone would be if he/she has no experience (*i.e.,* $E = e^*$). TE (*i.e.,* Eqn. (2)) measures the change of O when E changes from having no experience (*i.e.,* $E = e^*$) to having it (*i.e.,* $E = e$). Meanwhile, the Natural Direct Effect (NDE) and Total Indirect Effect (TIE) together make up TE [20] as follows,

$$\text{TE} = \text{NDE} + \text{TIE}. \tag{3}$$

Specifically, NDE denotes the effect of $E = e$ on O along the direct path $E \rightarrow O$ without considering the indirect effect along $E \rightarrow P \rightarrow O$. TIE measures the difference between TE and NDE, *i.e.,* the change of O when E changes from having no experience (*i.e.,* $E = e^*$) to having it (*i.e.,* $E = e$) along the indirect effect path $E \rightarrow P \rightarrow O$. For example, NDE of $E = e$ on O can be calculated by:

$$\text{NDE} = O_{e,p^*} - O_{e^*,p^*}, \tag{4}$$

where O_{e,p^*} denotes a counterfactual situation, E is set to e, and P is set to p^*, *i.e.,* the value of P under $E = e^*$ (see Figure 2(c)). Based on Eqn. (3) and (4), we can obtain TIE of $E = e$ on O by subtracting NDE from TE:

$$\text{TIE} = \text{TE} - \text{NDE} = O_{e,p} - O_{e,p^*}. \tag{5}$$

4 METHODOLOGY

We first formulate the task of OOD MSA, and then detail the proposed CLUE framework.

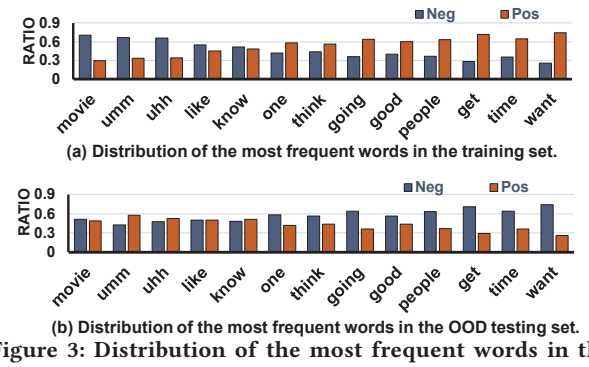


Figure 3: Distribution of the most frequent words in the training set and the OOD testing set of MOSEI. The y-axis shows the proportion of samples with the target word (*e.g.,* "movie") in the positive or negative categories.

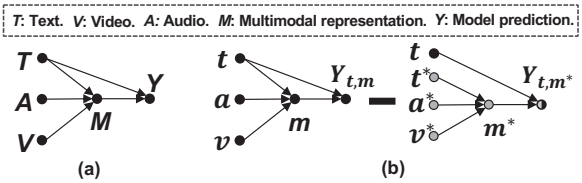


Figure 4: (a) The causal graph in the MSA. And (b) the illustration of counterfactual inference.

4.1 Task Formulation

Traditional MSA Task. A dataset with N training samples can be formulated as $\mathcal{D} = \{(t_1, a_1, v_1, y_1), \cdots, (t_N, a_N, v_N, y_N)\}$. Each quadruple (t_i, a_i, v_i, y_i) is associated with a video segment, whereby t_i, a_i, v_i, and y_i denotes the text, audio, video, and sentiment label of the i-th sample, respectively. The MSA aims to devise a multimodal model $\mathcal{F}_\theta$, which jointly considers the three modalities (*i.e.,* t_i, a_i, and v_i) as inputs and predicts the sentiment label (*i.e.,* y_i) as follows,

$$\hat{y}_i = \mathcal{F}_\theta(t_i, a_i, v_i), \tag{6}$$

where $\hat{y}_i$ denotes the predicted probability of sentiment labels for i-th sample and θ represents the learnable parameters.

OOD MSA Task. Traditional MSA models tend to heavily rely on textual modality for the sentiment analysis [8] because salient textual semantics (*e.g.,* "it is great.") easily reflect the sentiment tendency. However, there exist some spurious correlations between textual words and sentiment labels due to the selection bias [20] when collecting the dataset $\mathcal{D}$. For example, the word "movie" is highly correlated with the negative label in the MOSEI dataset as shown in Figure 3(a). The model trained with such biased data may prefer to classify the samples with "movie" as negative, which deteriorates the generalization performance of MSA models.

To reduce the influence of spurious correlations, we define an OOD MSA task, which aims to evaluate the OOD generalization performance of different MSA methods on the OOD testing set. Specifically, we develop an algorithm (illustrated in Section 5.1.3) to automatically construct the OOD testing set for each biased dataset, with significantly different word-sentiment correlations from the training one. The different distributions between the training and OOD testing sets can validate whether the MSA models are able to mitigate the effect of spurious correlations effectively.

4.2 CLUE Framework

We first inspect the causal relationships in the MSA task, and then detail the CLUE framework.

4.2.1 ***Causal Graph in the MSA Task.*** As shown in Figure 4(a), we utilize a causal graph to describe the causal relationships in the MSA task, *i.e.,* the relationships among the text T, audio A, video V, multimodal representation M, and model prediction Y. The explanations for the causal relationships are as follows:

- $T \rightarrow Y$ is a shortcut between the text and the model prediction because textual modality provides salient features with spurious correlations for the model prediction. This shortcut describes the bad effect of textual modality due to the spurious correlations between texts and labels.
- $(T, A, V) \rightarrow M \rightarrow Y$ denotes the indirect effect of T, A, and V on the model prediction Y via a mediator M. The multimodal representation M is obtained by the cross-modal modeling, which extracts more reliable textual semantics based on the alignment among different modalities. For example, the excited tone in the voice and the smiling face in the video guide M to focus on the word "great" in text and filter out the spurious correlations caused by "movie". Therefore, the indirect path $T \rightarrow M \rightarrow Y$ presents a good effect of texts.

Causal View of Traditional MSA Models. Traditional MSA models ignore the causal relationships in Figure 4, and thus blindly fit the correlations between the features of three modalities and sentiment labels as shown in Eqn. (6). As such, they mix the direct and indirect effects of textual modality on the model prediction, and then inevitably utilize the spurious correlations in the direct effect for unreliable predictions.

4.2.2 ***Counterfactual MSA Framework.*** Inspired by previous work [19, 31] embracing causal inference, we propose to model the causal relationships in Figure 4 based on the CLUE framework, which captures both the direct and indirect effects of textual modality. Following the notations in Section 3, the abstract format of causal relationships is:

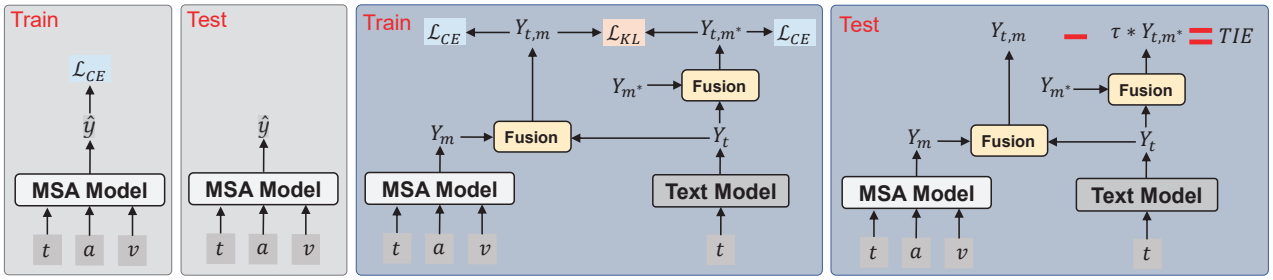
$$\begin{cases} Y_{t,m} = f_Y(T = t, M = m), \\ m = f_M(T = t, A = a, V = v). \end{cases} \tag{7}$$

where $Y_{t,m}$ denotes the model prediction under the situation of $M = m$ and $T = t$. Although Eqn. (7) provides the accurate formulation of causal relationships, $Y_{t,m}$ still suffers from the bad effect (*i.e.,* the direct effect) of $T = t$. This is because $Y_{t,m}$ contains both the direct and indirect effects. To disentangle them, we estimate NDE and TIE of textual modality on the model prediction separately. To accomplish this, we first calculate TE and NDE of texts, and then subtract NDE from TE to obtain TIE for the inference of CLUE.

Following the definition of causal effects in Section 3, TE on the model prediction Y under the situation of $T = t$, $A = a$ and $V = v$ can be calculated by

$$\text{TE} = Y_{t,m} - Y_{t^*,m^*} = f_Y(T = t, M = m) - f_Y(T = t^*, M = m^*), \tag{8}$$

where $m^* = f_M(T = t^*, A = a^*, V = v^*)$. Meanwhile, the symbols t^*, a^* and v^* denote the reference values of T, A, and V, respectively. In this work, we define the reference values as the situations where the features are not given, *i.e.,* using a dummy vector as the features of M. Next, we estimate the NDE of textual modality. Formally,



(a) Traditional MSA. (b) Counterfactual inference of CLUE.

Figure 5: Illustration of traditional MSA models and the counterfactual inference of CLUE.

the NDE of $T = t$ on the model prediction Y can be formulated as follows,

$$Y_{t,m^*} - Y_{t^*,m^*} = f_Y(T = t, M = m^*) - f_Y(T = t^*, M = m^*), \quad (9)$$

where T is set to t in the path $T \rightarrow Y$ while keeping $T = t^*$ along the path $T \rightarrow M \rightarrow Y$. $Y_{t,m^*} = f_Y(T = t, M = m^*)$ denotes the corresponding model prediction Y. Y_{t,m^*} answers a counterfactual problem: *what the model prediction would be if the MSA model only had seen the direct effect of the textual modality.* To mitigate the bad effect of $T = t$ in the direct path, in the inference stage, we subtract NDE from TE to calculate TIE of three modalities (see Figure 4(b)):

$$\text{TIE} = \text{TE} - \text{NDE} = Y_{t,m} - Y_{t,m^*}, \quad (10)$$

which mitigates the bad effect (*i.e.,* NDE) caused by the spurious correlations in the shortcut $T \rightarrow Y$. As such, we utilize the reliable indirect effect (*i.e.,* TIE) for the unbiased prediction by counterfactual inference.

Figure 1(d) illustrates the prediction result from a sample with the word "movie". In the prediction of NDE, this sample has a higher probability over the negative category due to the spurious correlation between words and the negative label. However, NDE is subtracted from TE, leading to a relatively higher probability over the positive category.

4.3 Instantiation of CLUE

To calculate TIE of textual modality, we implement the CLUE framework (see Figure 5) based on the causal graph in Figure 4(a).

Model Architecture. We implement two structural equations $f_Y(T = t, M = m)$ and $f_M(T = t, A = a, V = v)$ according to Eqn. (7). For $f_M(T = t, A = a, V = v)$, we utilize the cross-modal feature extractors in conventional MSA models, which take the text, audio, and video as inputs to calculate M. As to $f_Y(T = t, M = m)$, we implement it in a late-fusion manner inspired by existing research [19]. This manner can well utilize the strong representation capability of existing MSA models [31]. We thus reach the following,

$$f_Y(T = t, M = m) = h(Y_t, Y_m), \quad (11)$$

where $Y_t = g_Y(T = t)$ and $Y_m = g_Y(M = m) = g_Y(T = t, A = a, V = v)$ are the predicted probabilities by two neural models with different inputs, $g_Y(T = t)$ can be implemented by any traditional unimodal model, *e.g.,* BERT [2], $g_Y(T = t, A = a, V = v)$ can be instantiated by any multimodal MSA model, *e.g.,* MISA [8], and $h(\cdot)$ is a fusion function to combine the predictions Y_t and Y_m. In this work, we adopt the widely-adopted SUM fusion strategy:

$$Y_{t,m} = f_Y(T = t, M = m) = \text{SUM}(Y_t, Y_m) = \log \sigma(Y_t + Y_m), \quad (12)$$

where σ denotes the sigmoid function. In fact, since $Y_{t,m}$ represents the prediction probability and its sum should be one, we should apply the softmax operation after Eqn. (12). Meanwhile, whether using the softmax operation will not affect the prediction result, because we select the class with the maximal prediction score as the sentiment prediction.

Training. Towards optimization, we adopt the cross-entropy ($CE(\cdot)$) loss for $Y_{t,m}$ and Y_{t,m^*} as follows:

$$\mathcal{L}_{CE} = \alpha * \text{CE}(Y_{t,m}, y) + \beta * \text{CE}(Y_{t,m^*}, y), \quad (13)$$

where α and β are the weights to balance two loss functions, and y is the sentiment label for (t, a, v). In addition to using $Y_{t,m}$ to optimize TE, we also introduce a CE loss to facilitate the NDE prediction Y_{t,m^*} in the counterfactual world. We calculate $Y_{t,m^*} = h(Y_t, Y_{m^*})$ by Eqn. (12), where Y_{m^*} indicates the MSA model without (t, a, v) as inputs. Theoretically, we should utilize randomly initialized vectors or all-zero vectors to denote t^*, a^* and v^*, and then feed them into f_M to get m^*. Because only t^*, a^* and v^* affect the prediction of Y_{m^*} and Y_{m^*} is not affected by other features, we directly use a randomly initialized vector to represent the final result Y_{m^*} (i.e., the calculation result of t^*, a^* and v^*), and do not need to instantiate t^*, a^* and v^* in the real implementation. We thus use a trainable vector Y_{m^*} as the output of the MSA model, *i.e.,* a dummy vector shared for all samples.

An inappropriate Y_{m^*} may make the scales of TE and NDE different so that TIE would be dominated by one of them [19]. Therefore, we leverage Kullback-Leibler (KL) divergence to minimize the difference between Y_{t,m^*} and $Y_{t,m}$ to estimate Y_{m^*}:

$$\mathcal{L}_{KL} = \text{KL}(Y_{t,m^*}, Y_{t,m}). \quad (14)$$

To summarize, the final training loss is as follows:

$$\mathcal{L} = \sum_{(t_i, a_i, v_i, y_i) \in \mathcal{D}} \mathcal{L}_{CE} + \gamma * \mathcal{L}_{KL}, \quad (15)$$

where γ is the parameter to control the effect of $\mathcal{L}_{KL}$. We optimize CLUE in an end-to-end manner.

Inference. To leverage TIE for inference, the final prediction of CLUE can be obtained by,

$$\text{TIE} = Y_{t,m} - \tau * Y_{t,m^*} = h(Y_t, Y_m) - \tau * h(Y_t, Y_{m^*}), \quad (16)$$

where τ is the hyperparameter to control the degree of subtracting correlations between the textual modality and sentiment labels.

Algorithm 1 IID and OOD Set Construction.

Input: The whole dataset $\mathcal{D}$, the pre-defined distribution difference $\boldsymbol{\phi}_{\Delta}$, the number of iterations n, simulated annealing temperature τ, and the temperature decay rate α.
Output: IID set $\mathcal{D}_{iid}$ and OOD set $\mathcal{D}_{ood}$.

Get an IID set $\mathcal{D}_{iid}$ and OOD set $\mathcal{D}_{ood}$ by random splitting $\mathcal{D}$.
Compute distributions $\boldsymbol{\phi}_{iid}$ and $\boldsymbol{\phi}_{ood}$ of all words over different sentiment categories in $\mathcal{D}_{iid}$ and $\mathcal{D}_{ood}$, respectively.
Set $V = \|abs(\boldsymbol{\phi}_{iid} - \boldsymbol{\phi}_{ood}) - abs(\boldsymbol{\phi}_{\Delta})\|_1$.
repeat
  repeat
    Randomly swap samples between $\mathcal{D}_{iid}$ and $\mathcal{D}_{ood}$ by perturbation strategies[2] to a new IID set $\hat{\mathcal{D}}_{iid}$ and OOD set $\hat{\mathcal{D}}_{ood}$.
    Calculate $\hat{\boldsymbol{\phi}}_{iid}$ ($\hat{\boldsymbol{\phi}}_{ood}$) with $\hat{\mathcal{D}}_{iid}$ ($\hat{\mathcal{D}}_{ood}$), respectively.
    Set $\hat{V} = \|abs(\hat{\boldsymbol{\phi}}_{iid} - \hat{\boldsymbol{\phi}}_{ood}) - abs(\boldsymbol{\phi}_{\Delta})\|_1$.
    Get a random number R and $0 \leq R < 1$.
    if $V \geq \hat{V}$ **then**
      Set $\mathcal{D}_{iid}, \mathcal{D}_{ood}, V = \hat{\mathcal{D}}_{iid}, \hat{\mathcal{D}}_{ood}, \hat{V}$.
    else if $exp((\hat{V} - V)/\tau) > R$ **then**
      Set $\mathcal{D}_{iid}, \mathcal{D}_{ood}, V = \hat{\mathcal{D}}_{iid}, \hat{\mathcal{D}}_{ood}, \hat{V}$.
    end if
  until Swapping times reach n.
  Set $\tau = \tau * \alpha$.
until Iteration times reach n.

5 EXPERIMENTS

In this section, we conducted extensive experiments to answer the following research questions:

- **RQ1**. Does CLUE improve the performance of MSA models on the OOD testing?
- **RQ2**. How does CLUE perform on the biased testing, *i.e.,* the Independent and Identically Distributed (IID) testing?
- **RQ3**. How does each component of CLUE affect the results?

5.1 Experimental Settings

5.1.1 ***Datasets.*** For evaluation, we adopted two publicly accessible benchmark datasets on MSA.

MOSI [42] is a collection of 2,199 subjective utterance-video clips from 93 YouTube monologue videos. Each clip is labeled with a continuous sentiment score ranging from −3 (strongly negative) to 3 (strongly positive).

MOSEI [40] is an expanded version of MOSI with a larger number of utterance-video clips from more speakers and topics. Statistically, it comprises 23, 453 annotated utterance-video clips, derived from 5, 000 videos with 1, 000 distinct speakers and 250 different topics.

5.1.2 ***Evaluation Tasks and Metrics.*** Towards comprehensive evaluation, we assessed the MSA models with two widely-adopted tasks: 7-class classification and 2-class classification. The 7 sentiment classes in the former task range from strongly negative (−3) to strongly positive (3) and the accuracy (*Acc-7*) was adopted as the evaluation metric. Pertaining to the latter task, we further investigated two settings: *negative/non-negative* and *negative/positive*, as each has been once studied by previous work [29, 41]. For the *negative/non-negative* setting, we classified the sentiment of a given sample as the negative or non-negative class, while for the *negative/positive* setting, our goal is to assign each sample with a negative or positive sentiment. Towards the evaluation on the binary task, we used both accuracy (denoted as *Acc-2*) and $F1$ score.

5.1.3 ***IID and OOD Settings.*** For each task, we performed experiments in both IID and OOD settings. In the former setting, the testing set shares the IID distribution with the training one, which is similar to many research [9, 10, 15, 34, 35]; Whereas in the latter, we expected that the sample distribution of each word over different sentiment categories in the testing set is as different as possible from the training set. In a sense, for each task, it requires four sets of samples from each dataset: IID training, IID validation, IID testing, and OOD testing sets. Similar to previous IID studies, we still adopted the output of the MSA model to search the optimal parameters with the IID validation set. In the testing phase, we used TE of all modalities $Y_{t,m}$ without mitigating the bias for the biased IID testing, while the counterfactual inference results (*i.e.,* TIE in Eqn. (16)) for the OOD testing.

To split the datasets, we first divided each dataset (*i.e.,* MOSI or MOSEI) into two parts: the IID set, and the OOD set, where the former can be further split into three chunks (*i.e.,* IID training, IID validation, and IID testing) with random sampling, and the latter constitutes the OOD testing set. To achieve this, we adapted the simulated annealing algorithm, which is originally designed for solving the traveling salesman problem (TSP), for the dataset split. The adapted simulated annealing algorithm (as shown in Algorithm 1) will iterate until the overall distribution difference of all words over different sentiment categories between the IID set and OOD set gets sufficiently close to the pre-defined distribution difference $\boldsymbol{\phi}_{\Delta}$. The number of iterations is set to 800. The temperature parameter of this algorithm is set to 0.5, and the temperature decay is set to 0.99. Ultimately, the OOD testing set derived from MOSI contains 12 videos. To balance the size of OOD testing set and IID testing set, we randomly selected 12 videos from the IID set obtained by the above algorithm as the IID testing set, while the rest is divided into two groups: 59 (85%) for IID training and 10 (15%) for IID validation. Similarly, the IID training, IID validation, IID testing, OOD testing sets obtained from MOSEI contain 1, 830, 324, 330, and 330 videos, respectively. Towards intuitive understanding, we visualized the distributions of the frequent words in both the IID training set and the OOD testing set derived from MOSEI in Figure 3. As can be seen, their distributions are significantly different to each other.

5.1.4 ***Baselines.*** To evaluate the performance of CLUE, we adopted the following methods for comparison.

(1) **TFN**. The Tensor Fusion Network [39] learns both the intra-modality and inter-modality dynamics via Tensor Fusion and Modality Embedding Subnetworks, respectively.

(2) **LMF**. Low-rank Multimodal Fusion [16] performs multimodal fusion using modality-specific low-rank tensors, which decreases the computational complexity and improves the model's efficiency.

[2]Referring to the study [30], we utilized three perturbation strategies.

Table 1: OOD testing performance (%) comparison among different methods on MOSEI and MOSI datasets. For *Acc-2* and *F1*, "$*$" is calculated as "negative/non-negative" and "§" is calculated as "negative/positive". The best result of each pair of the original MSA model and the model with CLUE is highlighted in bold.

Model	MOSEI					MOSI				
	2-class				7-class	2-class				7-class
	Acc-2*	*F1**	*Acc-2*§	*F1*§	*Acc-7*	*Acc-2**	*F1**	*Acc-2*§	*F1*§	*Acc-7*
TFN	71.23	70.46	69.76	69.02	41.05	73.02	72.93	74.62	74.56	32.95
LMF	68.16	68.31	69.58	69.58	31.11	73.54	73.40	75.27	75.18	29.10
MulT	72.56	72.44	73.73	73.58	40.58	75.00	74.75	76.72	76.52	29.80
MAG-BERT	74.59	74.48	76.41	76.27	45.88	75.57	75.52	77.28	77.26	39.85
+CLUE (Ours)	**78.34**$^{+3.75}$	**78.23**$^{+3.75}$	**80.51**$^{+4.10}$	**80.46**$^{+4.19}$	**48.66**$^{+2.78}$	**77.25**$^{+1.68}$	**77.46**$^{+1.94}$	**78.65**$^{+1.37}$	**78.83**$^{+1.57}$	**40.75**$^{+0.90}$
MISA	74.48	74.39	76.45	76.33	43.15	75.90	75.82	77.39	77.35	38.05
+CLUE (Ours)	**77.17**$^{+2.69}$	**77.08**$^{+2.69}$	**78.77**$^{+2.32}$	**78.74**$^{+2.41}$	**46.86**$^{+3.71}$	**78.25**$^{+2.35}$	**78.28**$^{+2.46}$	**79.17**$^{+1.78}$	**79.19**$^{+1.84}$	**42.25**$^{+4.20}$
Self-MM	74.68	74.33	74.50	74.22	45.81	76.70	76.68	78.12	78.13	40.25
+CLUE (Ours)	**77.76**$^{+3.08}$	**77.72**$^{+3.39}$	**79.48**$^{+4.98}$	**79.47**$^{+5.25}$	**48.09**$^{+2.28}$	**78.75**$^{+2.05}$	**78.75**$^{+2.07}$	**79.94**$^{+1.82}$	**79.93**$^{+1.80}$	**41.75**$^{+1.50}$

Table 2: IID testing performance (%) comparison among different methods on MOSEI and MOSI datasets. *Acc-2* and *F1* are calculated as "negative/non-negative". We omitted the similar results of "negative/positive" to save space.

Model	MOSEI			MOSI		
	2-class		7-class	2-class		7-class
	Acc-2	*F1*	*Acc-7*	*Acc-2*	*F1*	*Acc-7*
TFN	81.59	81.54	52.11	80.24	80.31	40.07
LMF	79.59	80.34	48.27	79.85	79.95	35.04
MulT	81.05	81.44	53.21	79.61	79.71	35.19
MAG-BERT	82.82	83.19	53.52	83.91	83.96	46.97
+CLUE (Ours)	**84.62**	**85.46**	**53.68**	**84.37**	**84.28**	**48.84**
MISA	82.17	82.61	**53.26**	83.52	83.58	45.26
+CLUE (Ours)	**84.51**	**85.28**	53.15	**84.07**	**84.16**	**46.31**
Self-MM	83.71	83.80	53.31	84.14	84.17	**48.74**
+CLUE (Ours)	**84.52**	**84.46**	**53.42**	**84.31**	**84.38**	48.04

Table 3: Ablation study results (%) for the binary classification (negative/non-negative) of our proposed CLUE on MOSEI. The best results are highlighted in boldface.

Model	IID testing		OOD testing	
	Acc-2	*F1*	*Acc-2*	*F1*
MAG-BERT+CLUE	84.62	85.46	**78.34**	**78.23**
w/o-MSA model	80.31	81.52	65.49	68.01
w/o-text model	**85.09**	**85.60**	74.37	74.83
w/o-KL loss	84.55	84.45	78.28	78.17
MISA+CLUE	84.51	85.28	**77.17**	**77.08**
w/o-MSA model	80.31	81.52	65.49	68.01
w/o-text model	84.31	85.26	74.53	75.78
w/o-KL loss	**84.69**	**85.44**	76.75	76.77
Self-MM+CLUE	84.52	84.46	**77.76**	**77.72**
w/o-MSA model	80.31	81.52	65.49	68.01
w/o-text model	84.52	**85.41**	73.51	74.12
w/o-KL loss	**84.63**	84.53	77.41	77.43

(3) **MulT**. The core of Multimodal Transformer [29] is the cross-modal attention mechanism, which directly attends to low-level features in other modalities. These interactions lead to better performance on unaligned multimodal datasets.

(4) **MAG-BERT**. The Multimodal Adaptation Gate [25] regards the nonverbal behavior as a vector with a trajectory and magnitude, which is subsequently used to shift lexical representations within the pre-trained transformer model that demonstrates its advancement in many natural language tasks [2, 27].

(5) **MISA**. By projecting each modality of samples into two subspaces, this method learns both modality-invariant and -specific representations [8], which then are fused for sentiment analysis.

(6) **Self-MM**. This method introduces a unimodal label generation strategy based on the self-supervised method [37]. Furthermore, a novel weight self-adjusting strategy is introduced to balance different task loss constraints. It is worth mentioning that all the baselines treat the task of MSA as a regression task, which can obtain more supervisory information, while our CLUE treats MSA as a classification task.

5.1.5 ***Implementation Details.*** In this work, we adopted three state-of-the-art (SOTA) MSA models (*i.e.,* MISA, Self-MM, and MAG-BERT) as our MSA backbone of our CLUE framework, respectively. We use audio and video features provided by the MOSI and MOSEI followed the three MSA models. Regarding text modality, we also adopted the same setting with each of the three MSA models, where the pre-trained BERT is used and whether it should be fine-tuned depends on the original settings of these MSA models. Pertaining to the input of our Text Model branch, we simply adopted the 300D Glove [21] vectors as the word embeddings for our Text Model to better verify the effectiveness of our framework. Theoretically, the performance would be better if we use BERT embeddings. Regarding optimization, we adopted the Adam as the optimizer with the learning rate of $3e-5$ for our Text Model and learning rates of $6e-5$, $2e-5$, and $1e-5$ for different MSA backbones (*i.e.,* MISA, Self-MM, and MAG-BERT). The dropout rates for these three MSA backbones are set to 0.1, 0.3, and 0.1, respectively. The batchsize is set to 24 for MOSEI dataset, and 32 for MOSI dataset. We used the grid search strategy to find the optimal hyperparameters. Specifically, the hyperparameters α, β, γ are searched within $[0, 1]$ with the step size of 0.1, while the hyperparameter τ is searched within $[0, 2.5]$ with the corresponding step size of 0.15.

5.2 On Model Comparison (RQ1 & RQ2)

We adopted the SOTA MAG-BERT, MISA and Self-MM as the MSA backbone of our method, respectively. Table 1 and Table 2 show the performance comparison among different methods over OOD and IID testing settings, respectively. From these two tables, we have the following observations. 1) The models with our CLUE framework significantly outperform their original versions in the OOD testing setting. In particular, MAG-BERT+CLUE raises the *Acc-2* of MAG-BERT from 76.41% to 80.51% for negative/positive on MOSEI. This validates the superior generalization ability of our CLUE framework over existing methods. 2) Besides, even in the IID

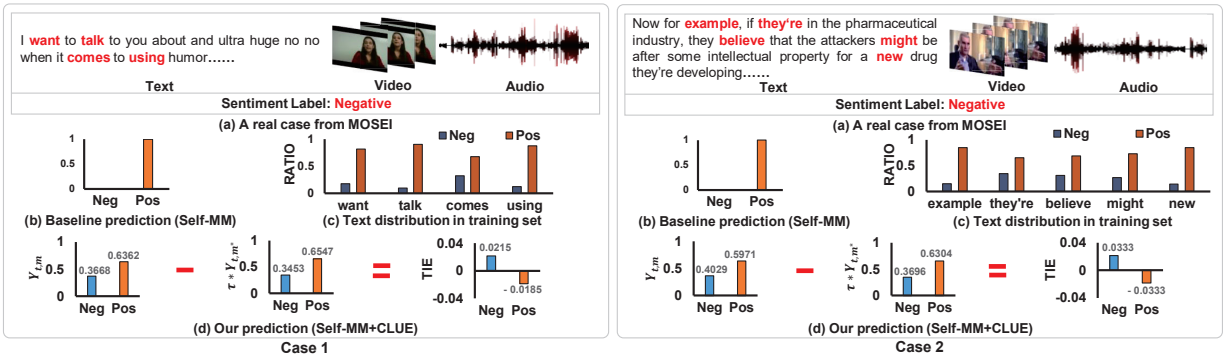


Figure 6: Cases of the binary sentiment classification (*i.e.,* positive (Pos) and negative (Neg) categories) by self-MM and CLUE. RATIO: the proportion of samples with the target words in the positive or negative categories of the training set.

testing, our MAG-BERT+CLUE also achieves the best performance in terms of most evaluation metrics on both datasets. And 3) for all the methods, their performance in the IID testing setting is better than that in the OOD testing setting. This confirms that the spurious correlations between textual words and sentiment labels do hurt the generalization ability of the model. However, we noticed that the performance decrease of our model is the least compared with that of the baseline methods. This reflects that our CLUE is more robust for the OOD testing.

5.3 On Ablation Study (RQ3)

To verify the effect of the MSA model branch, the text model branch, as well as the KL loss for regularizing $Y_{t,m}$ and Y_{t,m^*}, we introduced three derivatives of CLUE: **w/o-MSA model**, **w/o-text model**, and **w/o-KL loss**, by setting $\alpha = 0$ in Eqn. (13), $\beta = 0$ in Eqn. (13), and $\gamma = 0$ in Eqn. (15), respectively. Table 3 shows the ablation study results of our proposed CLUE on the MOSEI dataset with the task of binary classification. As can be seen, for IID testing, our MAG-BERT+CLUE, MISA+CLUE and Self-MM+CLUE achieve similar performance with their corresponding derivatives. This may be due to the fact that or because the bias is similar in training and testing set under IID setting. As for OOD testing, we have the following observations: 1) Our model largely outperforms the derivative w/o-MSA model with different MSA backbones (*i.e.,* MAG-BERT, MISA and Self-MM). This implies the necessity of introducing the multi-modal content of videos towards sentiment analysis. 2) For each backbone, our model exceeds the derivative w/o-text model across different evaluation metrics. This demonstrates the importance of mitigating the bad effect of textual modality with causal inference. And 3) both intact models surpass their w/o-KL loss derivatives. This shows that a proper Y_{m^*} is crucial for optimal performance.

5.4 On Case Study

To get an intuitive understanding of the generalization of our model, we illustrated the binary (*negative/positive*) classification results: Self-MM and our CLUE, which adopts Self-MM as the MSA model backbone, on two testing samples from MOSEI dataset in Figure 6. To gain deeper insights, we provided the distribution of the keywords mentioned in the sample's text, highlighted in red, over the positive and negative categories in the training set, as well as the classification results of each branch of our model. As can be seen, for both samples, our CLUE gives the correct sentiment, while Self-MM fails. Taking the Case 1 as an example, we can see that each branch of our model gives a higher probability to the positive category, so does the baseline Self-MM. However, combining the two branches with the causal inference, our model gives the negative prediction. This may be due to the fact that the keywords (*e.g.,* "example", "they're", and "believe") of the given testing sample in the training set are highly related to the positive sentiment. Thereby, the spurious correlation between words and sentiment labels mislead the two branches and the baseline Self-MM to predict the correlation rather than the causal relationship. This intuitively demonstrates the necessity of considering the spurious correlations between the text and sentiment. Similar observation can be found in Case 2.

6 CONCLUSION AND FUTURE WORK

In this work, we first analyzed the spurious correlations between textual modality and the sentiment labels, and then defined the OOD MSA task to evaluate the generalization ability of MSA models. Besides, we presented a model-agnostic counterfactual multimodal sentiment analysis framework that can retain the good effect of textual modality and mitigate its bad effect. Extensive experiments on the two benchmarks demonstrate the superiority of the proposed framework on both IID and OOD testing.

This work makes the initial step to consider OOD generalization in the MSA task, leaving many promising directions to future work. 1) As an initial work on the OOD MSA, we focus on the most harmful spurious correlations in textual modality. It is meaningful to explore the bad correlations in other modalities. 2) We adopt counterfactual reasoning to mitigate the direct effect of texts. To further improve the OOD generalization ability, another future direction is to discover more fine-grained causal relationships, *e.g.,* word-level correlation discovery. And 3) future work can study more strategies to implement the CLUE framework, such as superior fusion functions and text models.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.:U1936203; Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. 2013. Deep Canonical Correlation Analysis. In *Proceedings of the International Conference on Machine Learning*. JMLR.org, 1247–1255.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 4171–4186.

[3] Simon Dobrišek, Rok Gajšek, France Mihelič, Nikola Pavešić, and Vitomir Štruc. 2013. Towards efficient multi-modal emotion recognition. *International Journal of Advanced Robotic Systems* 10, 1 (2013), 53.

[4] Fuli Feng, Jizhi Zhang, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Empowering Language Understanding with Counterfactual Reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2226–2236.

[5] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[6] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

[7] Paridhi Gupta, Tanu Mehrotra, Ashita Bansal, and Bhawna Kumari. 2017. Multimodal sentiment analysis and context determination: Using perplexed Bayes classification. In *International Conference on Automation and Computing*. IEEE, 1–6.

[8] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1122–1131.

[9] Jun Hu, Shengsheng Qian, Quan Fang, Youze Wang, Quan Zhao, Huaiwen Zhang, and Changsheng Xu. 2021. Efficient graph deep learning in tensorflow with tf_geometric. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 3775–3778.

[10] Jun Hu, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2019. Hierarchical graph semantic pooling network for multi-modal community question answer matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1157–1165.

[11] Mahesh G Huddar, Sanjeev S Sannakki, and Vijay S Rajpurohit. 2018. An ensemble approach to utterance level multimodal sentiment analysis. In *International Conference on Computational Techniques, Electronics and Mechanical Systems*. IEEE, 145–150.

[12] Zaid Khan and Yun Fu. 2021. Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 3034–3042.

[13] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2928–2937.

[14] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. 2021. Interventional video relation detection. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 4091–4099.

[15] Fan Liu, Zhiyong Cheng, Lei Zhu, Zan Gao, and Liqiang Nie. 2021. Interest-Aware Message-Passing GCN for Recommendation. In *Proceedings of the Web Conference*. ACM, 1296–1305.

[16] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2247–2256.

[17] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. In *Proceedings of the Conference on Artificial Intelligence*. AAAI, 164–172.

[18] Hao Ni, Jingkuan Song, Xiaosu Zhu, Feng Zheng, and Lianli Gao. 2021. Camera-Agnostic Person Re-Identification via Adversarial Disentangling Learning. In *The ACM International Conference on Multimedia*. ACM, 2002–2010.

[19] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xiansheng Hua, and Jirong Wen. 2021. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 12700–12710.

[20] Judea Pearl. 2009. *Causality*. Cambridge university press.

[21] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1532–1543.

[22] Álvaro Peris and Francisco Casacuberta. 2019. A Neural, Interactive-predictive System for Multimodal Sequence to Sequence Tasks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 81–86.

[23] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. In *Proceedings of the Conference on Artificial Intelligence*. AAAI, 6892–6899.

[24] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual Inference for Text Classification Debiasing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 5434–5445.

[25] Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Mohammed E. Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2359–2369.

[26] J. Robins. 1986. A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Period. *Mathematical Modelling* 7, 9–12 (1986), 1393–1512.

[27] Xuemeng Song, Liqiang Jing, Dengtian Lin, Zhongzhou Zhao, Haiqing Chen, and Liqiang Nie. 2022. V2P: Vision-to-Prompt based Multi-Modal Product Summary Generation. In *The International Conference on Research and Development in Information Retrieval*. ACM, 992–1001.

[28] Teng Sun, Chun Wang, Xuemeng Song, Fuli Feng, and Liqiang Nie. 2022. Response Generation by Jointly Modeling Personalized Linguistic Styles and Emotions. *ACM Trans. Multim. Comput. Commun. Appl.* 18, 2 (2022), 52:1–52:20.

[29] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 6558–6569.

[30] Peter J. M. van Laarhoven and Emile H. L. Aarts. 1987. *Simulated Annealing: Theory and Applications*. Mathematics and Its Applications, Vol. 37. Springer. 7–15 pages.

[31] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. In *The International Conference on Research and Development in Information Retrieval*. ACM, 1288–1297.

[32] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. 2022. Causal Representation Learning for Out-of-Distribution Recommendation. In *Proceedings of the ACM Web Conference*. ACM, 3562–3571.

[33] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. TransModality: An End2End Fusion Method with Transformer for Multimodal Sentiment Analysis. In *The Web Conference*. ACM, 2514–2520.

[34] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 5382–5390.

[35] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1437–1445.

[36] Xu Yan, Li-Ming Zhao, and Bao-Liang Lu. 2021. Simplifying Multimodal Emotion Recognition with Single Eye Movement Modality. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1057–1063.

[37] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *Proceedings of the Conference on Artificial Intelligence*. AAAI, 10790–10797.

[38] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based Feature Reconstruction Network for Robust Multimodal Sentiment Analysis. In *The ACM International Conference on Multimedia*. ACM, 4400–4407.

[39] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1103–1114.

[40] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2236–2246.

[41] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention Recurrent Network for Human Communication Comprehension. In *Proceedings of the Conference on Artificial Intelligence*. AAAI, 5642–5649.

[42] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *IEEE Intelligent Systems* 36, 6 (2016), 82–88.

[43] Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Effective Sentiment-relevant Word Selection for Multi-modal Sentiment Analysis in Spoken Language. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 148–156.

[44] Xi Zhang, Feifei Zhang, and Changsheng Xu. 2021. Multi-Level Counterfactual Contrast for Visual Commonsense Reasoning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1793–1802.

[45] Lin Zheng, Naicheng Guo, Weihao Chen, Jin Yu, and Dazhi Jiang. 2020. Sentiment-guided Sequential Recommendation. In *The International Conference on Research and Development in Information Retrieval*. ACM, 1957–1960.