Debiased Multimodal Understanding for Human Language Sequences

Zhi Xu[1*], Dingkang Yang[1*†], Mingcheng Li[1], Yuzheng Wang[1], Zhaoyu Chen[1], Jiawei Chen[1], Jinjie Wei[1], Lihua Zhang[1,2,3‡]

[1]Academy for Engineering and Technology, Fudan University, Shanghai, China
[2]Institute of Metaverse & Intelligent Medicine, Fudan University, Shanghai, China
[3]Cognition and Intelligent Technology Laboratory, Fudan University, Shanghai, China
{zhixu20, dkyang20}@fudan.edu.cn

Abstract

Human multimodal language understanding (MLU) is an indispensable component of expression analysis (*e.g.*, sentiment or humor) from heterogeneous modalities, including visual postures, linguistic contents, and acoustic behaviours. Existing works invariably focus on designing sophisticated structures or fusion strategies to achieve impressive improvements. Unfortunately, they all suffer from the subject variation problem due to data distribution discrepancies among subjects. Concretely, MLU models are easily misled by distinct subjects with different expression customs and characteristics in the training data to learn subject-specific spurious correlations, limiting performance and generalizability across new subjects. Motivated by this observation, we introduce a recapitulative causal graph to formulate the MLU procedure and analyze the confounding effect of subjects. Then, we propose SuCI, a simple yet effective causal intervention module to disentangle the impact of subjects acting as unobserved confounders and achieve model training via true causal effects. As a plug-and-play component, SuCI can be widely applied to most methods that seek unbiased predictions. Comprehensive experiments on several MLU benchmarks clearly show the effectiveness of the proposed module.

Introduction

As a research hotspot that combines linguistic and nonverbal behaviours (*e.g.*, visual and acoustic modalities), human multimodal language understanding (MLU) has attracted significant attention from computer vision (Yang et al. 2022a), natural language processing (Tian et al. 2022), and speech recognition communities (Yang et al. 2022b,d) in recent years. Thanks to the progressive development of multimodal language benchmarks (Hasan et al. 2019; Zadeh and Pu 2018; Zadeh et al. 2016), extensive studies (Rahman et al. 2020; Han, Chen, and Poria 2021; Liu et al. 2018; Yu et al. 2021; Tsai et al. 2019; Liang et al. 2021b; Lv et al. 2021; Pham et al. 2019; Sun et al. 2022; Lei et al. 2023) have presented impressive multimodal models on training data containing distinct subjects, diverse topics, and different modalities. Despite the achievements of previous approaches by exploiting representation learning architectures (Yang et al. 2022a; Liang et al. 2021b) and fusion strategies (Tsai et al. 2019; Lv et al. 2021), they invariably suffer from a prediction bias when applied to testing samples of new subjects.

The harmful prediction bias is mainly caused by the subject variation problem. Specifically, different subjects' expression styles and behaviours (*e.g.*, facial expressions or acoustic information) from the training data are highly idiosyncratic in social communication, affected by the subjects' customs or culture (Li and Deng 2020). Once well-designed models are trained on such data, subject-specific semantic correlations (*e.g.*, particular facial action unit co-occurrence (Chen et al. 2022)) would inexorably affect performance and generalizability. Worse still, the spurious connections between the trained models and specific subjects will be transmitted via multiple modalities when the data paradigm is extended from isolated to multimodal situations. Recall the prominent MLU benchmarks (*e.g.*, MOSI (Zadeh et al. 2016) for sentiment analysis or UR_FUNNY (Hasan et al. 2019) for humor detection), whose collectors advocated video-level data splitting so that segments from the same video will not appear across train, valid, and test splits. Although the trained models may avoid memorizing the average affective state of a subject (Liang et al. 2021a), they cannot generalize well across new subjects. The examples in Figure 1 provide strong evidence of this. Concretely, subjects 1, 2, and 3 tend to use sentimentally unimportant words "*but*" and "*just*" to express negative emotions. In this case, the MLU model (Li, Wang, and Cui 2023) is misled to focus on spurious clues from the textual utterances of the subjects and make an entirely incorrect prediction when applied to subject 4. Similar observations are found in the visual and audio modalities. For instance, the trained model erroneously takes subject-specific facial appearances (*e.g.*, "grimace and pursed mouth" from subjects 1 and 2) and acoustic behaviours (*e.g.*, "agitated tone" from subjects 2 and 3) as semantic shortcuts to infer frustrating negative sentiment.

Motivated by the above observations, this paper aims to improve MLU methods by causal demystification rather than beating them. We propose a Subject Causal Intervention module (SuCI) to disentangle the impact of semantic differences among subjects from multimodal expressions. Specifically, we first formulate a universal causal graph to analyze the MLU procedure and interpret the causalities

*These authors contributed equally.
†Project lead.
‡Corresponding author.

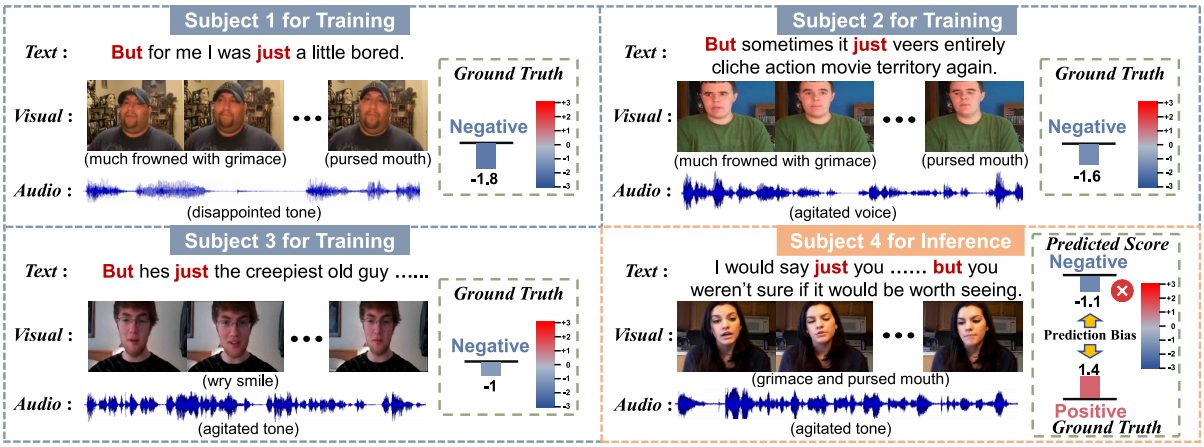

Figure 1: Examples on the MOSI benchmark illustrate the subject variation problem. Multimodal expressions from four subjects potentially convey distinct semantic correlations due to their different customs and styles in expressing sentiments.

among the variables. From the causal inference perspective, the *subjects* are essentially regarded as *confounders* (Glymour, Pearl, and Jewell 2016), which mislead the models to learn subject-specific semantic correlations in the training data, as well as causing prediction bias for new subjects in the inference phase. For clarity, we represent the multimodal inputs as $\boldsymbol{X}$ (cause) and its corresponding predictions as $\boldsymbol{Y}$ (effect). Existing MLU models aim to approximate $P(\boldsymbol{Y}|\boldsymbol{X})$ as much as possible while failing to perform unbiased predictions. Unlike conventional likelihood estimation $P(\boldsymbol{Y}|\boldsymbol{X})$, our SuCI achieves subject de-confounded training and removes subject-caused confounding effects by embracing causal intervention $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ regarding the backdoor adjustment rule (Pearl 2009a). As a model-agnostic component, SuCI can be readily integrated into most MLU models to perform unbiased predictions by pursuing true causal effects among variables.

To sum up, this paper has the following contributions. **(i)** We investigate the subject variation problem in MLU tasks via a tailored causal graph and identify the subjects as confounders which misleads the models to capture subject-specific spurious correlations and cause prediction bias. **(ii)** Based on the causal theory of backdoor adjustment, we present SuCI, a subject causal intervention module to remove prediction bias and confounding effects of subjects. **(iii)** We evaluate the effectiveness of SuCI on several MLU benchmarks. Experimental results show that SuCI can significantly and consistently improve existing baselines, achieving new SOTA performance.

Related Work

Human Multimodal Language Understanding. Benefiting from available human communication resources and data, MLU benchmarks (Zadeh and Pu 2018; Zadeh et al. 2016; Hasan et al. 2019) with different scales and typologies have been increasingly developed and applied in recent years. Recent MLU tasks focus on subject-centered intention understanding and behavior analysis from text, visual, and audio modalities, including but not limited to emotion recognition (Lv et al. 2021), sentiment analysis (Han, Chen, and Poria 2021), and humor detection (Hasan et al. 2019). Considering the heterogeneous nature of multimodal languages, numerous works have presented seminal network structures (Pham et al. 2019; Liang et al. 2021b; Yu et al. 2021; Sun et al. 2022), fusion strategies (Tsai et al. 2019, 2018; Zadeh et al. 2017; Rahman et al. 2020), and representation learning paradigms (Yang et al. 2022a,d,c; Li, Yang, and Zhang 2023). For instance, MFSA (Yang et al. 2022d) presented a factorized representation strategy to learn similarities and differences among multimodal languages. Despite achievements, existing methods invariably suffer from performance bottlenecks due to subject-related prediction bias. In comparison, we identify subjects in MLU benchmarks as harmful confounders from a causal perspective and improve different models with our SuCI.

Causal Demystification. Causal demystification as a potential statistical theory aims to pursue causal effects among observed variables rather than their shallow correlations. Benefiting from the advances in learning-based technologies (Yang et al. 2023d,a,c,b; Liu et al. 2023a,b), several studies exploring causality are mainly divided into two channels: causal intervention (Lin et al. 2022; Yang et al. 2021) and counterfactual reasoning (Qian et al. 2021; Tang et al. 2020). Intervention (Pearl 2009a) focuses on altering the natural tendency of the independent variable to vary with other variables to eliminate the impact of adverse effects. Counterfactuals depict imagined outcomes produced by factual variables under different treatment conditions (Pearl 2009b). In this paper, we address the confounding effect from multiple modalities by embracing causal intervention, which is more adaptable for multimodal approaches.

Methodology

Structural Causal Graph in MLU Tasks

To systematically diagnose the confounding effect present in MLU tasks, we first design a recapitulative causal graph to summarize the causal relationships among variables.

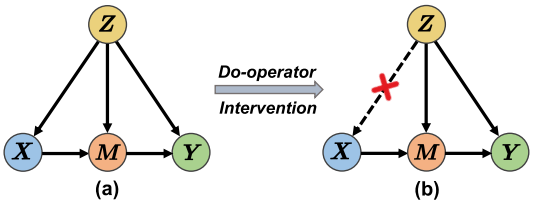

Figure 2: The causal graph explains causal effects of MLU procedure. Nodes denote variables and arrows denote the direct causal effects. (a) The conventional likelihood estimation $P(\boldsymbol{Y}|\boldsymbol{X})$. (b) The causal intervention $P(\boldsymbol{Y}|do(\boldsymbol{X}))$.

The mainstream graphical notation following the structured causal model (Pearl et al. 2000) is adopted due to its intuitiveness and interpretability. Concretely, a causal graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ is considered a directed acyclic graph, which represents how a set of variables $\mathcal{N}$ convey causal effects through the causal links $\mathcal{E}$. As shown in Figure 2(a), there are four variables in the MLU causal graph, including multimodal inputs $\boldsymbol{X}$, multimodal representations $\boldsymbol{M}$, unintended confounders $\boldsymbol{Z}$, and predictions $\boldsymbol{Y}$. Note that our causal graph applies to various MLU approaches since it is highly general, imposing no constraints on the detailed implementations. The causalities are shown below.

► **Link $\boldsymbol{Z} \rightarrow \boldsymbol{X}$.** Multimodal expressions from distinct subjects are recorded to produce multimodal inputs $\boldsymbol{X}$, where $\boldsymbol{X}$ is a generalized definition of text $\boldsymbol{X}_t$, visual $\boldsymbol{X}_v$, and audio $\boldsymbol{X}_a$ modalities for simplicity, *i.e.*, $\boldsymbol{X} = \{\boldsymbol{X}_t, \boldsymbol{X}_v, \boldsymbol{X}_a\}$. Subjects are identified as harmful confounders $\boldsymbol{Z}$ due to the subject-related prediction bias caused by different human expression customs and differences (Hasan et al. 2019; Zadeh and Pu 2018). In this case, $\boldsymbol{Z}$ is a collective denomination of multimodal confounding sources. For training inputs $\boldsymbol{X}$, $\boldsymbol{Z}$ determines the subject-related biased content that is recorded, *i.e.*, $\boldsymbol{Z} \rightarrow \boldsymbol{X}$.

► **Link $\boldsymbol{Z} \rightarrow \boldsymbol{M} \leftarrow \boldsymbol{X}$.** $\boldsymbol{M}$ denotes the refined multimodal representations extracted by any MLU model, which acts as a mediator before the final classifier. The link $\boldsymbol{Z} \rightarrow \boldsymbol{M}$ indicates detrimental $\boldsymbol{Z}$ confounding models to capture subject-specific characteristics embedded in $\boldsymbol{M}$ to produce spurious semantic correlations. Several intuitive examples are the semantically unimportant words from the text modality and the agitated tones from the audio modality in Figure 1. Furthermore, $\boldsymbol{M}$ contains universal multimodal feature semantics from $\boldsymbol{X}$ that can be reflected via the causal link $\boldsymbol{X} \rightarrow \boldsymbol{M}$.

► **Link $\boldsymbol{M} \rightarrow \boldsymbol{Y} \leftarrow \boldsymbol{Z}$.** The causal path $\boldsymbol{M} \rightarrow \boldsymbol{Y}$ reveals that the impure $\boldsymbol{M}$ confounded by $\boldsymbol{Z}$ further impacts the final predictions $\boldsymbol{Y}$ of downstream MLU tasks. Meanwhile, adverse confounders' prior information in the training data implicitly interferes with $\boldsymbol{Y}$ along the link $\boldsymbol{Z} \rightarrow \boldsymbol{Y}$.

According to the causal theory (Pearl 2009a), the confounders $\boldsymbol{Z}$ are the common cause of $\boldsymbol{X}$ and corresponding predictions $\boldsymbol{Y}$. The positive effect of the subject-agnostic multimodal semantics provided by $\boldsymbol{M}$ follows the desired causal path $\boldsymbol{X} \rightarrow \boldsymbol{M} \rightarrow \boldsymbol{Y}$, which we aim to achieve and pursue. Unfortunately, $\boldsymbol{Z}$ causes subject-related prediction bias and misleads trained models to learn subject-specific misleading semantics rather than pure causal effects, leading to biased predictions on uninitiated subjects. The detrimental effects follow the backdoor paths $\boldsymbol{X} \leftarrow \boldsymbol{Z} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{X} \leftarrow \boldsymbol{Z} \rightarrow \boldsymbol{M} \rightarrow \boldsymbol{Y}$.

Causal Intervention via Backdoor Adjustment

Following the causal graph in Figure 2(a), the MLU model relies on the likelihood $P(\boldsymbol{Y}|\boldsymbol{X})$ for predictions that suffer from backdoor effects, which can be decomposed by the Bayes rule as follows:

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{\boldsymbol{z}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{M} = \mathcal{F}_m(\boldsymbol{X}, \boldsymbol{z}))P(\boldsymbol{z}|\boldsymbol{X}), \quad (1)$$

where $\mathcal{F}_m(\cdot)$ denotes any vanilla MLU model to learn the multimodal representations $\boldsymbol{M}$. $\boldsymbol{z}$ is a stratum of confounders (*i.e.*, a subject), which introduces the observational bias via $P(\boldsymbol{z}|\boldsymbol{X})$. Theoretically, an ideal solution would be to collect massive data samples to ensure that subjects with different expression characteristics are included in the training and testing sets. However, this way is unrealistic due to social ethics issues (Jones 1999). To address this, we embrace causal intervention $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ to interrupt the adverse effects propagating between $\boldsymbol{X}$ and $\boldsymbol{Y}$ along the backdoor paths via the backdoor adjustment theory (Pearl 2009a). The $do(\cdot)$ operator is an efficient approximation to implement the empirical intervention (Glymour, Pearl, and Jewell 2016). In our case, backdoor adjustment means measuring the causal effect of each stratum in the subject confounders and then performing a weighted integration based on the prior proportions of samples from different subjects in the training data to estimate the average causal effect. From Figure 2(b), the impact from $\boldsymbol{Z}$ to $\boldsymbol{X}$ is cut off since the model would enable the subject prototype as the confounder in each stratum to contribute equally to the predictions $\boldsymbol{Y}$ by $P(\boldsymbol{Y}|do(\boldsymbol{X}))$. Eq. (1) with the intervention is formulated as:

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) = \sum_{\boldsymbol{z}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{M} = \mathcal{F}_m(\boldsymbol{X}, \boldsymbol{z}))P(\boldsymbol{z}). \quad (2)$$

The model is no longer disrupted by subject-specific spurious correlations in backdoor paths since $\boldsymbol{z}$ no longer affects $\boldsymbol{X}$. $P(\boldsymbol{z})$ is the prior probability that depicts the proportion of each $\boldsymbol{z}$ in the whole.

Subject De-confounded Training with SuCI

We present a plug-in Subject Causal Intervention module (SuCI) to convert the theoretical intervention in Eq. (2) into a practical implementation. As Figure 3 shows, SuCI can be readily integrated into the vanilla MLU model to estimate $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ through the subject de-confounded training. The implementation details are as follows.

Subject-specific Feature Disentanglement. How to effectively disentangle subject-specific features from heterogeneous modalities is key to determining confounders. Considering that subject-related spurious semantics are usually distributed among different frames of multimodal sequences (Yang et al. 2022a), we first devise a dynamic fusion mechanism to aggregate the frame-level semantic clues. Let

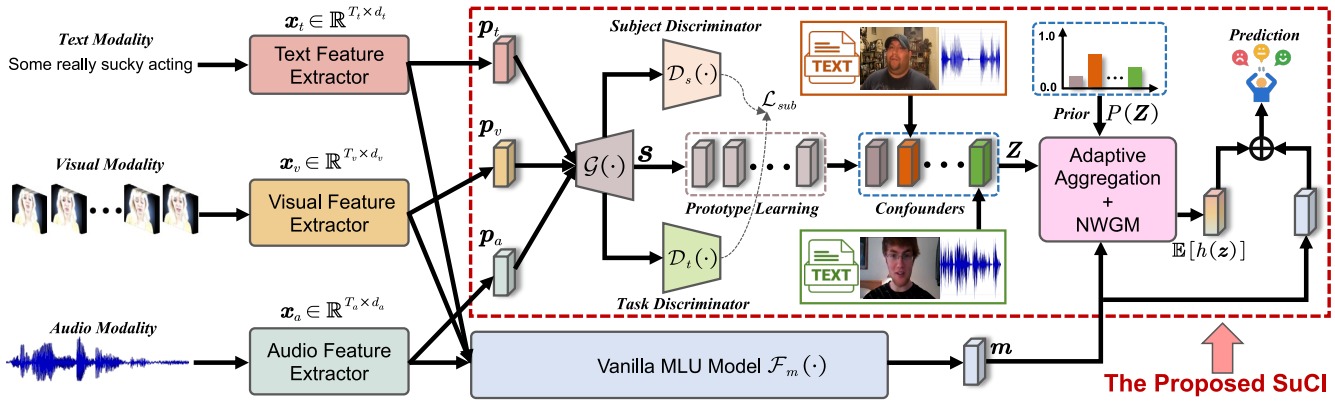

Figure 3: A general MLU pipeline for the subject de-confounded training. The red dotted box shows the core component that achieves the approximation to causal intervention: our SuCI. SuCI can be readily integrated into the vanilla MLU model via backdoor adjustment to mitigate subject-specific spurious correlations and achieve debiased predictions in downstream tasks.

the text, visual, and audio modality sequences from the corresponding feature extractors be denoted as $\boldsymbol{x}_t \in \mathbb{R}^{T_t \times d_t}$, $\boldsymbol{x}_v \in \mathbb{R}^{T_v \times d_v}$, and $\boldsymbol{x}_a \in \mathbb{R}^{T_a \times d_a}$, respectively, where $T_{(\cdot)}$ is the frame length and $d_{(\cdot)}$ is the embedding dimension. The dynamic fusion procedure is formulated as follows:

$$\xi_m = \phi(\boldsymbol{x}_m \boldsymbol{w}_{x_m} + \boldsymbol{b}_{x_m}) \in \mathbb{R}^{T_m \times 1}, \tag{3}$$

$$\boldsymbol{p}_m = \xi_m^T \boldsymbol{x}_m \in \mathbb{R}^{d_m}, \tag{4}$$

where $m \in \{t, v, a\}$, $\phi(\cdot)$ is the softmax function, $\boldsymbol{w}_{x_m} \in \mathbb{R}^{d_m \times 1}$, and $\boldsymbol{b}_{x_m} \in \mathbb{R}^{T_m \times 1}$ are the learnable parameters. The attention vectors ξ_m adaptively capture salient semantics and produce informative representations $\boldsymbol{p}_m$ based on the contributions of different frames. Subsequently, we introduce an adversarial strategy to disentangle the subject-specific semantics and avoid the task-related semantics. The design philosophy is to distill pure subject features for better confounder construction. Specifically, a subject generator $\mathcal{G}(\cdot;\theta_\mathcal{G})$ is presented to project $\boldsymbol{p}_t$, $\boldsymbol{p}_v$, and $\boldsymbol{p}_a$ of multimodal utterances from the subject into a common space to yield a subject-specific feature $\boldsymbol{s}$, expressed as follows:

$$\boldsymbol{s} = \mathcal{G}([\boldsymbol{p}_t, \boldsymbol{p}_v, \boldsymbol{p}_a]; \theta_\mathcal{G}) \in \mathbb{R}^{d_s \times 1}, \tag{5}$$

where $[,]$ stands for the concatenation operator and $d_s = d_t + d_v + d_a$, The generator is implemented as a two-layer perceptron with GeLU (Hendrycks and Gimpel 2016) activation. $\boldsymbol{s}$ is distilled by the following objective:

$$\mathcal{L}_{sub} = \mathcal{CE}(\mathcal{D}_s(\boldsymbol{s};\theta_{\mathcal{D}_s}), y_s) + \mathcal{MSE}(\mathcal{D}_t(\boldsymbol{s};\theta_{\mathcal{D}_t}), \frac{1}{C_t}), \tag{6}$$

where $\mathcal{CE}(\cdot)$ and $\mathcal{MSE}(\cdot)$ represent the cross-entropy loss and mean squared error loss, respectively. $\mathcal{D}_s(\cdot;\theta_{\mathcal{D}_s})$ and $\mathcal{D}_t(\cdot;\theta_{\mathcal{D}_t})$ are the subject discriminator and task discriminator parameterized by $\theta_{\mathcal{D}_s}$ and $\theta_{\mathcal{D}_t}$, respectively, which consist of feed-forward neural layers. y_s is the subject label determined by the dataset's subject index. C_t is the number of categories in downstream tasks. In Eq. (6), $\boldsymbol{s}$ is supervised to encourage the output logits of the task discriminator to be equally distributed among all prediction categories to exclude task-related semantic information. Also, the subject discriminator is optimized to ensure that $\boldsymbol{s}$ belongs to a given subject, *i.e.*, contains subject-specific semantics.

Confounder Construction. Confounder construction aims to make the model measure the causal effect of subject confounders among different strata during training to avoid subject-related prediction bias. Considering that subject features are similar within the same stratum and different across strata (Pearl 2009a), we utilize all the features from a specific subject as the subject prototype, which represents the universal attributes of confounders in a specific stratum. Concretely, we maintain a memory cache for each subject during training to store and update the subject prototype as a confounder, which is computed as $\boldsymbol{z}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} \boldsymbol{s}_k^i$. N_i is the number of training samples for the i-th subject, and $\boldsymbol{s}_k^i$ denotes the k-th feature of the i-th subject. $\boldsymbol{z}_i$ is updated at the end of each epoch. In practice, each $\boldsymbol{z}_i$ is initialized by the uniform distribution to ensure stable training at the first epoch. Based on the number of subjects N_c, a stratified confounder dictionary is constructed, which is formulated as $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{N_c}]$.

Intervention Instantiation. Estimating $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ in practice is high overhead since it requires the forward computation of $\boldsymbol{X}$ and $\boldsymbol{z}$ for each pair. To reduce the overhead, we introduce the Normalized Weighted Geometric Mean (NWGM) (Xu et al. 2015) to achieve feature-level approximation for intervention instantiation:

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) \overset{\text{NWGM}}{\approx} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{M} = \sum_{\boldsymbol{z}} \mathcal{F}_m(\boldsymbol{X}, \boldsymbol{z}) p(\boldsymbol{z})). \tag{7}$$

In this case, causal intervention makes $\boldsymbol{X}$ contribute fairly to the predictions of $\boldsymbol{Y}$ by incorporating every $\boldsymbol{z}$. We parameterize a network model to approximate Eq. (7) as follows:

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) = \boldsymbol{W}_m \boldsymbol{m} + \boldsymbol{W}_h \mathbb{E}[h(\boldsymbol{z})], \tag{8}$$

where $\boldsymbol{W}_m \in \mathbb{R}^{d_h \times d}$ and $\boldsymbol{W}_h \in \mathbb{R}^{d_h \times d_s}$ are the learnable parameters. $\boldsymbol{m} \in \mathbb{R}^{d \times 1}$ is the multimodal representation produced by the vanilla MLU model. The above approximation is reasonable since the effect on $\boldsymbol{Y}$ is attributed to $\boldsymbol{M}$ and $\boldsymbol{Z}$ in the MLU causal graph. Then, $\mathbb{E}[h(\boldsymbol{z})]$ is approximated as an adaptive aggregation for all confounders

Methods	$Acc_7 \uparrow (\%)$	$Acc_2 \uparrow (\%)$	$F1 \uparrow (\%)$
EF-LSTM	33.7	75.3	75.2
LF-LSTM	35.3	76.8	76.7
Graph-MFN	29.6	75.4	76.6
TFN	32.1	73.9	73.4
LMF	32.8	76.4	75.7
MFM	36.2	78.1	78.1
RAVEN	33.2	78.0	76.6
MCTN	35.6	79.3	79.1
TCSP	-	80.9	81.0
PMR	40.6	83.6	83.4
FDMER	42.1	84.2	83.9
MulT	39.5	82.6	82.5
MulT + **SuCI**	**40.7**$^{+1.2}$	**83.4**$^{+0.8}$	**83.4**$^{+0.9}$
MISA	40.2	81.8	81.9
MISA + **SuCI**	**41.6**$^{+1.4}$	**83.3**$^{+1.5}$	**83.1**$^{+1.2}$
Self-MM	41.6	83.9	84.1
Self-MM + **SuCI**	**42.0**$^{+0.4}$	**84.3**$^{+0.4}$	**84.6**$^{+0.5}$
MMIM	41.8	84.1	84.1
MMIM + **SuCI**	**42.4**$^{+0.6}$	**84.9**$^{+0.8}$	**84.8**$^{+0.7}$
DMD	41.0	83.3	83.2
DMD + **SuCI**	**42.2**$^{+1.2}$	**84.6**$^{+1.3}$	**84.5**$^{+1.3}$

Table 1: Comparison results of different methods and SuCI-based models on the MOSI benchmark. Improved results and corresponding gains compared to vanilla methods are marked in **bold** and red, respectively.

Methods	$Acc_7 \uparrow (\%)$	$Acc_2 \uparrow (\%)$	$F1 \uparrow (\%)$
EF-LSTM	47.4	78.2	77.9
LF-LSTM	48.8	80.6	80.6
Graph-MFN	45.0	76.9	77.0
RAVEN	50.0	79.1	79.5
MCTN	49.6	79.8	80.6
TCSP	-	82.8	82.7
PMR	52.5	83.3	82.6
FDMER	53.8	83.9	83.8
MulT	51.2	82.1	81.9
MulT + **SuCI**	**52.4**$^{+1.2}$	**83.2**$^{+1.1}$	**82.9**$^{+1.0}$
MISA	51.3	82.3	82.3
MISA + **SuCI**	**52.6**$^{+1.3}$	**83.5**$^{+1.2}$	**83.2**$^{+0.9}$
Self-MM	52.9	83.9	83.8
Self-MM + **SuCI**	**53.6**$^{+0.7}$	**84.2**$^{+0.3}$	**84.2**$^{+0.4}$
MMIM	53.7	84.4	84.3
MMIM + **SuCI**	**54.4**$^{+0.7}$	**85.5**$^{+1.1}$	**85.2**$^{+0.9}$
DMD	53.5	84.1	84.0
DMD + **SuCI**	**54.6**$^{+1.1}$	**85.8**$^{+1.7}$	**85.7**$^{+1.7}$

Table 2: Comparison results of different models on the MOSEI benchmark.

according to backdoor adjustment, which is formulated as:

$$\mathbb{E}[h(\boldsymbol{z})] = \sum_{i=1}^{N_c} \psi_i \boldsymbol{z}_i p(\boldsymbol{z}_i), \tag{9}$$

$$\psi_i = \phi(\frac{(\boldsymbol{W}_q \boldsymbol{m})^T (\boldsymbol{W}_k \boldsymbol{z}_i)}{\sqrt{d_s}}), \tag{10}$$

where $\boldsymbol{W}_q \in \mathbb{R}^{d_n \times d}$ and $\boldsymbol{W}_k \in \mathbb{R}^{d_n \times d_s}$ are mapping matrices. $p(\boldsymbol{z}_i) = \frac{N_i}{N}$, where N is the number of training samples. In practice, $\boldsymbol{m}$ from one sample queries each $\boldsymbol{z}_i$ in the confounder dictionary $\boldsymbol{Z} \in \mathbb{R}^{N_c \times d_s}$ to obtain the sample-specific attention set $\{\psi_i\}_{i=1}^{N_c}$. The intuition insight is that samples from one subject are impacted by varying degrees of confounder $\boldsymbol{z}_i$ of other subjects.

Objective Optimization. The standard cross-entropy loss is employed as the task-related training objective, which is expressed as $\mathcal{L}_{task} = -\frac{1}{N_i} \sum_{i=1}^{N_i} y_t \cdot log\hat{y}_t$, where y_t is the task-related ground truth. The overall objective function is computed as $\mathcal{L}_{all} = \mathcal{L}_{sub} + \mathcal{L}_{task}$.

Experiments

Benchmarks and Model Zoo

Benchmarks. Extensive experiments are conducted on three mainstream MLU benchmarks. Concretely, **MOSI** (Zadeh et al. 2016) is a multimodal human sentiment analysis dataset consisting of 2,199 video segments. The standard data partitioning is 1,284 samples for training, 284 samples for validation, and 686 samples for testing. These samples contain a total of 89 distinct subjects from video blogs. Each sample is manually annotated with a sentiment score ranging from -3 to 3. **MOSEI** (Zadeh and Pu 2018) is a large-scale human sentiment and emotion recognition benchmark containing 22,856 video clips from 1,000 different subjects and 250 diverse topics. Among these samples, 16,326, 1,871, and 4,659 samples are used as training, validation and testing sets. The sample annotation protocols used for sentiment analysis are consistent with MOSI. **UR_FUNNY** (Hasan et al. 2019) is a multimodal human humor detection dataset that contains 16,514 video clips from 1,741 subjects collected by the TED portal. There are 10,598, 2,626, and 3,290 samples in the training, validation, and testing sets. Each sample provides the target punchlines and associated contexts from multimodal utterances to support the detection of subject humor/non-humor in the binary labels. Following widely adopted implementations (Yang et al. 2022a; Li, Wang, and Cui 2023), we use the 7-class accuracy (Acc_7), binary accuracy (Acc_2), and $F1$ score to evaluate the results on MOSI and MOSEI. The standard binary accuracy (Acc_2) is utilized for evaluation on UR_FUNNY.

Model Zoo. Considering the applicability, we choose five representative models with different network structures to verify the proposed plug-in SuCI. A brief overview is as follows. **MulT** (Tsai et al. 2019) constructs multimodal transformers to learn element dependencies between pairwise modalities. **MISA** (Devamanyu, Roger, and Soujanya 2020) captures modality-invariant and modality-specific diverse representations based on feature disentanglement. **Self-MM** (Yu et al. 2021) utilizes a self-supervised paradigm to learn additional emotion semantics from unimodal label generation. **MMIM** (Han, Chen, and Poria 2021) proposes a hierarchical mutual information maximization to alleviate the loss of valuable task-related clues. **DMD** (Li, Wang, and Cui 2023) designs cross-modal knowledge distillations to bridge the semantic gap among modalities.

Implementation Details

All models follow consistent feature extraction procedures for fair comparisons. The text feature extractor is instantiated by pre-trained Glove word embedding tool (Pennington, Socher, and Manning 2014) to obtain 300-dimensional linguistic vectors. For MOSI & MOSEI, we use the library Facet (iMotions 2017) to extract an ensemble of visual features, including 35 facial action units to reflect facial ges-

Methods	Context	Punchline	$Acc_2 \uparrow (\%)$
C-MFN	✓		58.45
C-MFN		✓	64.47
TFN		✓	64.71
LMF		✓	65.16
C-MFN	✓	✓	65.23
FDMER		✓	70.55
MulT		✓	66.65
MulT + **SuCI**		✓	**67.88**$^{+1.23}$
MISA		✓	67.34
MISA + **SuCI**		✓	**68.96**$^{+1.62}$
Self-MM		✓	68.77
Self-MM + **SuCI**		✓	**69.72**$^{+0.95}$
MMIM		✓	69.53
MMIM + **SuCI**		✓	**70.92**$^{+1.39}$
DMD		✓	68.70
DMD + **SuCI**		✓	**70.84**$^{+2.14}$

Table 3: Comparison results of different models on the UR_FUNNY benchmark.

ture changes. Meanwhile, Openface (Baltrušaitis, Robinson, and Morency 2016) is utilized on UR_FUNNY to extract 75-dimensional features related to facial behaviors and expressions. The audio feature extraction is executed utilizing the software COVAREP (Degottex et al. 2014) to obtain diverse acoustic attributes, where the dimensions on MOSI & MOSEI and UR_FUNNY are 74 and 81, respectively. The word-aligned data points are employed across all benchmarks. In the SuCI implementation, the hidden dimensions d_h and d_n are set to 64 and 128, respectively. The size d_s of each subject confounder is 325, 409, and 456 on MOSI, MOSEI, and UR_FUNNY, respectively. We implement the selected methods and SuCI on NVIDIA Tesla A800 GPUs utilizing the PyTorch toolbox, where other training settings are aligned to their original protocols.

Comparison with State-of-the-art Methods

We compare the SuCI-based models with extensive SOTA methods, including EF-LSTM, LF-LSTM, C-MFN (Hasan et al. 2019), Graph-MFN (Zadeh and Pu 2018), TFN (Zadeh et al. 2017), MFM (Tsai et al. 2018), RAVEN (Wang et al. 2019), MCTN (Pham et al. 2019), TCSP (Wu et al. 2021), PMR (Lv et al. 2021), and FDMER (Yang et al. 2022a).

Results on MOSI Benchmark. (i) From Table 1, SuCI consistently improves the performance of the selected methods on all metrics. Concretely, the overall gains of Acc_7, Acc_2, and $F1$ scores among the five models increased by 4.8%, 4.8%, and 4.6%, respectively. These improvements confirm that our module can break through the performance bottlenecks of most baselines in a model-agnostic manner. **(ii)** SuCI can bring more favorable gains for decoupled learning-based efforts. For instance, MISA and DMD achieve salient relative improvements among different metrics of 1.8%∼3.5% and 1.6%∼2.9%, respectively. A reasonable explanation is that the decoupling pattern diffuses the spurious semantics caused by subject confounders in multiple feature subspaces. In this case, SuCI's de-confounding ability is more effective. **(iii)** Compared to recent PMR and FDMER with complex network stacking and massive parameters (Lv et al. 2021; Yang et al. 2022a), MMIM achieves the best results by equipping our SuCI.

Methods	$Acc_7 \uparrow (\%)$	$Acc_2 \uparrow (\%)$	$F1 \uparrow (\%)$
MulT	46.9	77.6	77.4
MulT + **SuCI**	**48.5**$^{+1.6}$	**80.2**$^{+2.6}$	**80.3**$^{+2.9}$
MISA	46.6	77.2	77.1
MISA + **SuCI**	**48.3**$^{+1.7}$	**78.9**$^{+1.7}$	**79.4**$^{+2.3}$
Self-MM	47.7	78.5	78.3
Self-MM + **SuCI**	**48.2**$^{+0.5}$	**79.5**$^{+1.0}$	**79.6**$^{+1.3}$
MMIM	49.3	79.6	79.3
MMIM + **SuCI**	**51.8**$^{+2.5}$	**82.5**$^{+2.9}$	**82.4**$^{+3.1}$
DMD	48.9	80.8	80.7
DMD + **SuCI**	**51.6**$^{+2.7}$	**82.4**$^{+1.6}$	**82.4**$^{+1.7}$

Table 4: Cross-dataset evaluation of models trained on the MOSI training set and tested on the MOSEI testing set.

Methods	$Acc_7 \uparrow (\%)$	$Acc_2 \uparrow (\%)$	$F1 \uparrow (\%)$
MulT	37.4	80.2	79.9
MulT + **SuCI**	**38.7**$^{+1.3}$	**81.7**$^{+1.5}$	**81.5**$^{+1.6}$
MISA	37.8	80.5	80.7
MISA + **SuCI**	**39.0**$^{+1.2}$	**82.2**$^{+1.7}$	**82.3**$^{+1.6}$
Self-MM	38.9	82.0	82.1
Self-MM + **SuCI**	**39.5**$^{+0.6}$	**82.7**$^{+0.7}$	**82.4**$^{+0.3}$
MMIM	39.3	82.5	82.5
MMIM + **SuCI**	**41.1**$^{+1.8}$	**83.4**$^{+0.9}$	**83.3**$^{+0.8}$
DMD	38.6	81.6	81.4
DMD + **SuCI**	**40.6**$^{+2.0}$	**83.0**$^{+1.4}$	**82.9**$^{+1.5}$

Table 5: Cross-dataset evaluation of models trained on the MOSEI training set and tested on the MOSI testing set.

Results on MOSEI Benchmark. Table 2 provides performance comparison results on MOSEI. **(i)** The SuCI-based models outperform the vanilla methods by large margins on all metrics. For example, SuCI helps the latest DMD to achieve new SOTA performance with considerable absolute gains of 1.1%, 1.7%, and 1.7% in Acc_7, Acc_2, and $F1$ scores, respectively. These observations show the broad applicability and usefulness of our module in removing the subject-related prediction bias. **(ii)** The improvements provided by SuCI on MOSEI are more significant than those on MOSI. The plausible deduction is that MOSEI contains richer subjects and their highly idiosyncratic multimodal utterances in various scenarios. Thus, SuCI can more accurately remove spurious correlations caused by appropriately extracted subject confounders and offer sufficient gains.

Results on UR_FUNNY Benchmark. From Table 3, we show the detection results using the target punchline for fair comparisons with other works. **(i)** The SuCI-based models achieve consistent performance increases and accomplish competitive and better results than previous methods. These substantial improvements imply the necessity of performing subject de-confounded training in detecting human humor. **(ii)** In particular, MMIM and DMD equipped with SuCI yield absolute gains of 1.39% and 2.14%, achieving new SOTAs with the Acc_2 of 70.92% and 70.84%, respectively.

Cross-dataset Evaluation

Since the used MOSI and MOSEI benchmarks have the same annotation protocols, we establish cross-dataset evaluations for MOSI-training→MOSEI-testing and MOSEI-training→MOSI-testing in Tables 4 and 5, respectively. The

Setting	MOSI		MOSEI		UR_FUNNY	
	MISA	DMD	MISA	DMD	MISA	DMD
Vanilla Method	81.8	83.3	82.3	84.1	67.34	68.70
+ **The Proposed SuCI**	**83.3**	**84.6**	**83.5**	**85.8**	**68.96**	**70.84**
Necessity of Subject Feature Learning						
w/o DFM	82.9	84.2	83.2	85.3	68.64	70.67
w/o Subject Discriminator	82.2	83.8	82.7	85.0	68.57	69.69
w/o Task Discriminator	82.6	84.1	82.9	85.3	68.62	70.44
Importance of Different Modalities						
w/o Text Modality	82.4	83.6	82.7	84.6	68.25	69.38
w/o Visual Modality	82.9	84.3	83.1	85.4	68.62	69.94
w/o Audio Modality	82.7	84.1	83.0	85.1	68.77	70.35
Rationality of Confounder Dictionary						
w/ Random $\boldsymbol{Z}$	79.3	81.0	78.2	79.9	62.46	84.05
w/ Clustered $\boldsymbol{Z}$	82.8	83.7	82.9	85.0	67.98	69.63
Effectiveness of Adaptive Aggregation						
w/o ψ_i	82.7	84.2	83.1	85.4	68.45	70.32
w/o $p(\boldsymbol{z}_i)$	83.0	84.5	83.4	85.6	68.73	70.66

Table 6: Ablation studies on the three benchmarks. "w/" and "w/o" mean the with and without. "DFM" is the Dynamic Fusion Mechanism.

design intuition is that exploring prediction performance on testing data with different distributions than the training data (*i.e.*, out-of-distribution, OOD) helps verify confounding effects and model generalizability. The five vanilla methods show severe performance deterioration compared to the results in the Independent Identically Distributed (IID) setting from Tables 1 and 2. For instance, the testing results on MOSEI and MOSI decreased by the average Acc_7 of 4.64% and 2.42% across all methods. This is inevitable since spurious correlations between trained models and specific subjects are exacerbated and amplified in uninitiated subjects under the OOD setting. Fortunately, our SuCI significantly improves the results of all models in cross-dataset evaluations. These substantial gains confirm that our module favorably mitigates the subject variation problem and enhances the generalizability and robustness of the vanilla models.

Ablation Studies

In Table 6, we perform ablation studies to evaluate the impact of all components in SuCI. We report results for the Acc_2 metrics due to similar trends for the other metrics.

Necessity of Subject Feature Learning. **(i)** We first replace our dynamic fusion mechanism with the average pooling operation along the frame length to obtain refined representations $\boldsymbol{p}_m$. The insufficient gains on all benchmarks suggest that assigning adaptive weights based on the distinct frame element contributions in multimodal sequences facilitates better capturing subject-related semantics. **(ii)** Subsequently, we separately remove the subject and task discriminators to investigate the role of feature disentanglement. Intuitively, the subject discriminator provides substantial gains for both methods since it supervises the generator to yield discriminative subject features $\boldsymbol{s}$ that are better used for confounder construction. **(iii)** The task discriminator is equally critical as it ensures that task-related information is excluded from $\boldsymbol{s}$ to distill the pure subject bias.

Importance of Different Modalities. **(i)** When the text, visual, and audio modalities from the subjects are removed separately in de-confounded training, the improvements in SuCI for the baselines show significant deterioration. These decreased results confirm that subject-specific spurious characteristics are transmitted in multimodal utterances and that considering the complete modalities is necessary. **(ii)** The reason for the pronounced effect of the text modality may be the adverse statistical shortcuts caused by the highly uneven distribution of textual words across the samples of distinct subjects, which is a confounder inducer.

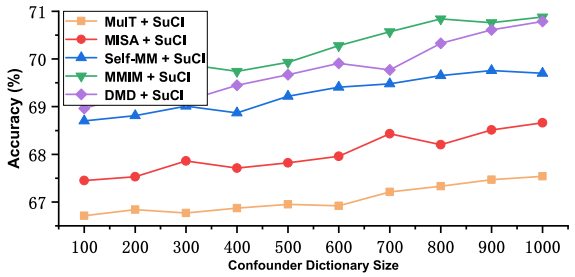

Figure 4: Ablation study results for the number of subject confounders on the UR_FUNNY benchmark.

Rationality of Confounder Dictionary. We provide two candidates of the same size as the default confounder dictionary $\boldsymbol{Z}$ to evaluate the rationality of our confounder construction. These two alternative dictionaries are obtained by random initialization and unsupervised K-Means++ clustering (Bahmani et al. 2012). **(i)** The random $\boldsymbol{Z}$ would significantly impair the performance and even underperform the vanilla methods, justifying the proposed subject prototypes as confounders. **(ii)** The performance gains of the clustered $\boldsymbol{Z}$ are sub-optimal due to the lack of subject-specific information supervision, leading to the indistinguishability and coupling of confounding effects across different subjects.

Effectiveness of Adaptive Aggregation. Here, the adaptive aggregation strategy is explored by eliminating the attention weights ψ_i and the prior probabilities $p(\boldsymbol{z}_i)$ in $\mathbb{E}[h(\boldsymbol{z})]$, respectively. The consistent performance drops on all benchmarks imply that characterizing the importance and proportion of each subject confounder is indispensable for achieving effective causal intervention based on subject debiasing.

Impact of Subject Confounder Number. Finally, we test the impact of different numbers of subject confounders on SuCI performance in Figure 4. The SuCI-based models all exhibit overall rising gain trends as the training subjects increase. These findings suggest that sufficient stratified confounders facilitate better backdoor adjustment implementation and accurate average causal effect estimation.

Conclusion

This paper is the first to reveal the long-neglected subject variation problem in MLU tasks and identify subjects as essentially harmful confounders from a causal perspective. Thus, we present a subject causal intervention module (SuCI) to remove the prediction bias caused by the subject-specific spurious correlations. Extensive experiments show the broad applicability of SuCI in diverse methods.

Acknowledgments

This work is supported in part by the National Key R&D Program of China (2021ZD0113503).

References

Bahmani, B.; Moseley, B.; Vattani, A.; Kumar, R.; and Vassilvitskii, S. 2012. Scalable k-means++. *arXiv preprint arXiv:1203.6402*.

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. IEEE.

Chen, Y.; Chen, D.; Wang, T.; Wang, Y.; and Liang, Y. 2022. Causal intervention for subject-deconfounded facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 374–382.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 960–964. IEEE.

Devamanyu, H.; Roger, Z.; and Soujanya, P. 2020. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, volume 34, 1122–1131.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Han, W.; Chen, H.; and Poria, S. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Hasan, M. K.; Rahman, W.; Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; et al. 2019. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

iMotions. 2017. *Facial expression analysis*.

Jones, B. D. 1999. Bounded rationality. *Annual review of political science*, 2(1): 297–321.

Lei, Y.; Yang, D.; Li, M.; Wang, S.; Chen, J.; and Zhang, L. 2023. Text-oriented Modality Reinforcement Network for Multimodal Sentiment Analysis from Unaligned Multimodal Sequences. *arXiv preprint arXiv:2307.13205*.

Li, M.; Yang, D.; and Zhang, L. 2023. Towards Robust Multimodal Sentiment Analysis under Uncertain Signal Missing. *IEEE Signal Processing Letters*, 30: 1497–1501.

Li, S.; and Deng, W. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*.

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6631–6640.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M. A.; Zhu, Y.; et al. 2021a. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*.

Liang, T.; Lin, G.; Feng, L.; Zhang, Y.; and Lv, F. 2021b. Attention Is Not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8148–8156.

Lin, X.; Wu, Z.; Chen, G.; Li, G.; and Yu, Y. 2022. A causal debiasing framework for unsupervised salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 1610–1619.

Liu, Y.; Yang, D.; Fang, G.; Wang, Y.; Wei, D.; Zhao, M.; Cheng, K.; Liu, J.; and Song, L. 2023a. Stochastic video normality network for abnormal event detection in surveillance videos. *Knowledge-Based Systems*, 110986.

Liu, Y.; Yang, D.; Wang, Y.; Liu, J.; and Song, L. 2023b. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *arXiv preprint arXiv:2302.05087*.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Lv, F.; Chen, X.; Huang, Y.; Duan, L.; and Lin, G. 2021. Progressive Modality Reinforcement for Human Multimodal Emotion Recognition From Unaligned Multimodal Sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2554–2562.

Pearl, J. 2009a. Causal inference in statistics: An overview. *Statistics Surveys*, 3: 96–146.

Pearl, J. 2009b. *Causality*. Cambridge University Press.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19: 2.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6892–6899.

Qian, C.; Feng, F.; Wen, L.; Ma, C.; and Xie, P. 2021. Counterfactual Inference for Text Classification Debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5434–5445.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Conference Association for Computational Linguistics Meeting (ACL)*, volume 2020, 2359. NIH Public Access.

Sun, H.; Wang, H.; Liu, J.; Chen, Y.-W.; and Lin, L. 2022. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 3722–3729.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3716–3725.

Tian, B.; Cao, Y.; Zhang, Y.; and Xing, C. 2022. Debiasing NLU models via causal intervention and counterfactual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 11376–11384.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Conference. Association for Computational Linguistics Meeting (ACL)*, volume 2019, 6558. NIH Public Access.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, 7216–7223.

Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; and Zhu, L.-N. 2021. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics Meeting (ACL-IJCNLP)*, 4730–4738.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2048–2057. PMLR.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022a. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 1642–1651.

Yang, D.; Huang, S.; Liu, Y.; and Zhang, L. 2022b. Contextual and Cross-modal Interaction for Multi-modal Speech Emotion Recognition. *IEEE Signal Processing Letters*, 1–5.

Yang, D.; Huang, S.; Wang, S.; Liu, Y.; Zhai, P.; Su, L.; Li, M.; and Zhang, L. 2022c. Emotion Recognition for Multiple Context Awareness. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13697, 144–162.

Yang, D.; Huang, S.; Xu, Z.; Li, Z.; Wang, S.; Li, M.; Wang, Y.; Liu, Y.; Yang, K.; Chen, Z.; Wang, Y.; Liu, J.; Zhang, P.; Zhai, P.; and Zhang, L. 2023a. AIDE: A Vision-Driven Multi-View, Multi-Modal, Multi-Tasking Dataset for Assistive Driving Perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 20459–20470.

Yang, D.; Kuang, H.; Huang, S.; and Zhang, L. 2022d. Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 1708–1717.

Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; and Zhang, L. 2023b. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*.

Yang, K.; Yang, D.; Zhang, J.; Li, M.; Liu, Y.; Liu, J.; Wang, H.; Sun, P.; and Song, L. 2023c. Spatio-Temporal Domain Awareness for Multi-Agent Collaborative Perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 23383–23392.

Yang, K.; Yang, D.; Zhang, J.; Wang, H.; Sun, P.; and Song, L. 2023d. What2comm: Towards Communication-Efficient Collaborative Perception via Feature Decoupling. In *Proceedings of the 31th ACM International Conference on Multimedia (ACM MM)*, 7686–7695.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9847–9857.

Yu, W.; Xu, H.; Yuan, Z.; and 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 (AAAI)*, volume 35, 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zadeh, A.; and Pu, P. 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 (ACL)*.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.