applied sciences

MDPI

Article

Causal Inference for Modality Debiasing in Multimodal Emotion Recognition

Juyeon Kim †, Juyoung Hong † and Yukyung Choi *

Department of Convergence Engineering for Intelligent Drone, Sejong University, Gwangjin-gu, Seoul 05006, Republic of Korea; jykim@rcv.sejong.ac.kr (J.K.); jyhong@rcv.sejong.ac.kr (J.H.)
* Correspondence: ykchoi@sejong.ac.kr
† These authors contributed equally to this work.

Abstract: Multimodal emotion recognition (MER) aims to enhance the understanding of human emotions by integrating visual, auditory, and textual modalities. However, previous MER approaches often depend on a dominant modality rather than considering all modalities, leading to poor generalization. To address this, we propose Causal Inference in Multimodal Emotion Recognition (CausalMER), which leverages counterfactual reasoning and causal graphs to capture relationships between modalities and reduce direct modality effects contributing to bias. This allows CausalMER to make unbiased predictions while being easily applied to existing MER methods in a model-agnostic manner, without requiring any architectural modifications. We evaluate CausalMER on the IEMOCAP and CMU-MOSEI datasets, widely used benchmarks in MER, and compare it with existing methods. On the IEMOCAP dataset with the MulT backbone, CausalMER achieves an average accuracy of 83.4%. On the CMU-MOSEI dataset, the average accuracies with MulT, PMR, and DMD backbones are 50.1%, 48.8%, and 48.8%, respectively. Experimental results demonstrate that CausalMER is robust in missing modality scenarios, as shown by its low standard deviation in performance drop gaps. Additionally, we evaluate modality contributions and show that CausalMER achieves balanced contributions from each modality, effectively mitigating direct biases from individual modalities.

Keywords: emotion recognition; multimodal learning; causal inference

Citation: Kim, J.; Hong, J.; Choi, Y. Causal Inference for Modality Debiasing in Multimodal Emotion Recognition. *Appl. Sci.* **2024**, *14*, 11397. https://doi.org/10.3390/app142311397

Academic Editor: Jing Jin

Received: 25 October 2024
Revised: 26 November 2024
Accepted: 5 December 2024
Published: 6 December 2024

1. Introduction

Emotion recognition is essential for enabling more natural interactions between humans and machines [1]. It has been widely applied in fields such as Human–Computer Interaction (HCI), robotics [2], and healthcare [3], where recognizing emotions enhances user experiences [4]. Recently, multimodal emotion recognition (MER) has attracted considerable attention as an approach to improving both the accuracy and sophistication of emotion recognition systems. Multimodal emotion recognition involves the integration of various data types to identify emotions. This approach is inspired by the observation that humans naturally rely on multiple cues—visual, auditory, and textual—when interpreting emotions [5]. Utilizing multiple modalities offers rich information that can significantly enhance the understanding of human emotions and intentions.

A key challenge in multimodal emotion recognition lies in developing fusion strategies to produce meaningful representations. However, in certain multimodal scenarios, one dominant modality may outperform the others and hinder the optimal integration of the remaining modalities [6]. This type of problem, where the model becomes biased toward one modality, is referred to as modality bias [7,8]. It can disrupt the overall optimization of the recognition system, preventing it from fully harnessing the complementary strengths of each modality. Furthermore, in real-world scenarios, not all modalities are consistently available [3,6]. For instance, text data might be missing due to speech recognition errors, or audio data may be inaccessible because of a faulty sensor. In such cases, models that

heavily rely on a dominant modality struggle to adapt, limiting their ability to handle incomplete or missing data effectively. As shown in the first column of Figure 1, this often leads to poor generalization, as the model becomes less capable of adapting to variations in input or missing modalities [6].

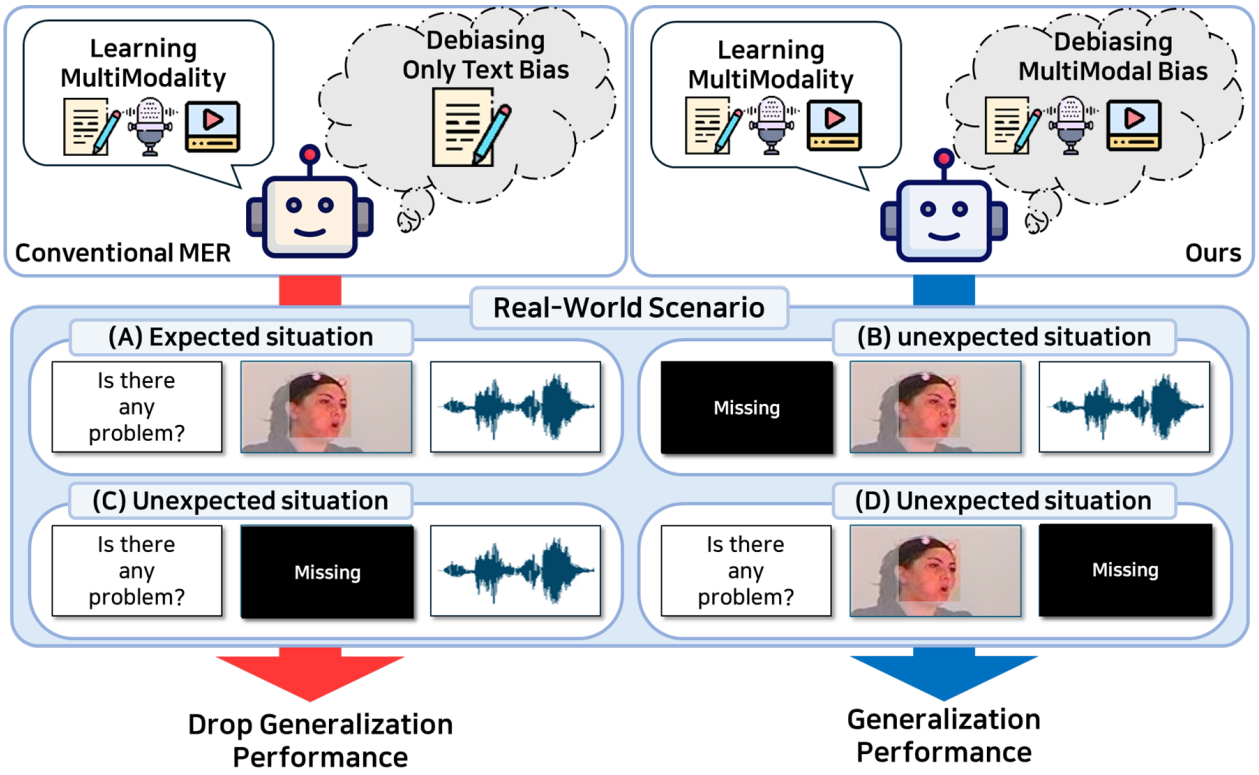

Figure 1. Problem Definition. If the model is debiased considering only language, it cannot effectively handle modality-biased samples where audio and video data are missing, as it fails to address biases in the other modalities.

Many multimodal models tend to rely heavily on the language modality [6]. Previous works [9–11] have addressed the issue of modality bias in multimodal learning, particularly focusing on the role of the language modality. While these methods propose to mitigate language modality bias, they often overlook biases present in other modalities, such as video and audio data. In many multimodal datasets, the importance of a particular modality can vary at the sample level and is not always dominant in every situation [12]. Existing methods effectively handle language bias but struggle to address biases in audio or video modalities. Additionally, these methods may encounter difficulties when certain modalities, like video or audio, are missing or imbalanced in the input data. Therefore, it is crucial to account for biases in all modalities—language, video, and audio—when designing multimodal systems to ensure robust performance across various scenarios.

By leveraging counterfactual reasoning, we can disentangle various effects in multimodal learning, enabling effective modality debiasing. Motivated by causal inference and counterfactual reasoning [13,14], we propose a novel approach called **Causal** Inference in **M**ultimodal **E**motion **R**ecognition (CausalMER) to reduce bias in text, audio, and video modalities. CausalMER can be easily applied to existing MER methods in a model-agnostic manner without requiring any architectural modifications. We construct a causal graph to capture the relationships between modalities, allowing us to understand how different modalities influence the training process. Using this causal graph, CausalMER estimates the direct modality effects caused by modality bias. During inference, CausalMER calculates the total indirect effect by subtracting the direct modality effects from the total causal effect,

effectively eliminating direct modality bias. Furthermore, experimentation with the MER backbone and CausalMER in the context of modality absence revealed that CausalMER demonstrates robustness to absent modalities.

The contributions of this work can be summarized as follows:

- We propose CausalMER, a novel approach based on causal inference, to reduce three-modality (text, audio, and video) bias in multimodal emotion recognition.
- We determined greater robustness in scenarios of modality absence compared to MER backbones.
- CausalMER was designed in a model-agnostic fashion. Thus, we can utilize any MER backbones.

2. Related Works

2.1. Multimodal Emotion Recognition

Multimodal emotion recognition aims to infer human emotions by integrating data from textual, visual, and auditory sources. The heterogeneity between these modalities provides diverse and complementary information, enriching the understanding of emotional states. Consequently, previous multimodal methods have concentrated on designing sophisticated fusion strategies to combine these different modalities. The early method [15] simply used feature-level fusion by concatenating features from different modalities. Another type of fusion is the attention mechanism [16], which considers the relative importance of each modality. Since the introduction of the transformer [17], cross-modal attention-based fusion methods [18–21] have significantly advanced, leading to more robust and enhanced modality representations. Existing methods focus on effective fusion, often combining modalities into a joint embedding space that overlooks the uniqueness of each modality [22]. Therefore, recent methods [22–24] aim to learn separate representations for each modality to reduce information redundancy in multimodal data. A progressive reinforcement mechanism [23] is used to learn potential adaptive relationships between different modalities from multimodality to unimodal. The feature disentanglement-based method [22] aims to learn modality-specific and modality-invariant subspaces for effective multimodal fusion. The graph-based distillation method [24] enhances modality-specific features, allowing for adaptive cross-modal knowledge transfer.

However, most previous work in multimodal emotion recognition has primarily focused on how to fuse different modalities. Without addressing modality bias, these approaches struggle to handle situations where one or more modalities are missing.

2.2. Counterfactual Reasoning

Counterfactual reasoning is a tool for examining causal effects in specific situations. It involves depicting imagined outcomes by applying factual variables in alternative treatment scenarios [25]. By reevaluating the causal relationship between preceding events and outcomes via counterfactual reasoning, one can rectify erroneous causal connections or reasoning fallacies. That is why it serves as a debiasing methodology across multiple tasks, including emotion recognition [26], sentiment classification [10,14], fake news detection [9], and visual question answering (VQA) [13]. Counterfactual reasoning is occasionally employed to eliminate a particular bias in the dataset within a unimodal context [26], and it is also utilized to detect and mitigate multiple biases [14]. Furthermore, counterfactual reasoning is applicable not only to unimodal scenarios but also to multimodal situations, primarily to mitigate biases inherent in datasets, as demonstrated in CLUE [10] and MCIS [11]. Most present approaches employ counterfactual reasoning to identify biases inherent in the dataset. Conversely, CF-VQA [13] has, for the first time, tackled the modality bias present in multimodal learning using counterfactual reasoning methods. In Visual Question Answering (VQA), models predominantly emphasize text over images to produce answers, and CF-VQA [13] addressed and enhanced this textual modality bias through counterfactual reasoning. Consequently, approaches to address modality bias arising during model training in multimodal contexts via counterfactual

reasoning have been examined; nevertheless, the majority focus on mitigating particular biases inherent in the dataset.

In multimodal scenarios, the majority of studies demonstrate modality bias, and in scenarios involving three or more modalities, one or more modality biases may arise. In a multimodal scenario, neglecting the biases of all modalities may result in unintended biases emerging in modalities other than the one from which the bias was eliminated. The absence of the relied-upon modality in a real-world context can result in a substantial drop in performance. This paper presents CausalMER, which eliminates biases across all modalities via counterfactual reasoning to tackle this phenomenon.

3. Preliminaries

In this section, we introduce the key concepts and notations related to causal inference, referencing the work [13]. We denote random variables with capital letters and their observed values with lowercase letters.

Causal graph. A causal graph is a probabilistic graphical model that reflects the causal relationships among variables. It is represented as a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N}$ denotes the set of variables and $\mathcal{E}$ represents the corresponding causal relationships. Figure 2a illustrates an example of a causal graph with three variables. The causality from cause variable X to effect variable Y consists of two parts: the direct effect (i.e., $X \to Y$) and the indirect effect, which occurs through the mediator variable M (i.e., $X \to M \to Y$). For example, a causal graph illustrating the effect of medication on disease treatment can be represented as follows: $X = x$ refers to the medication, Y represents the disease treatment outcome, and M includes other factors like the patient's age, physical condition, and other relevant variables.

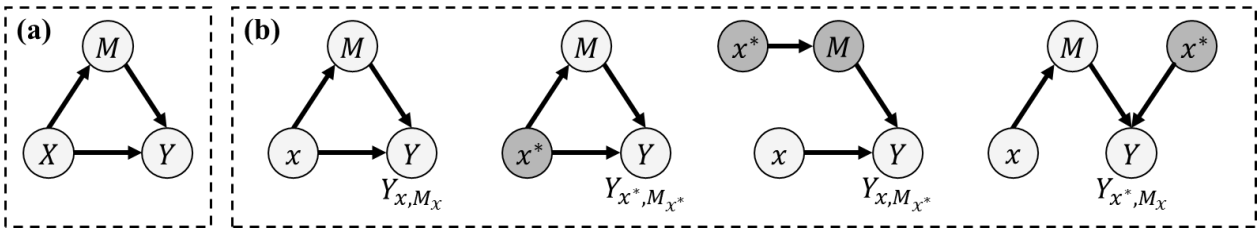

Figure 2. Causal graph. (**a**) Example of a causal graph. (**b**) Examples of counterfactual notations. White nodes represent the value $X = x$, while gray nodes correspond to the value $X = x^*$. Here, x^* indicates a counterfactual scenario.

Counterfactual inference. Counterfactual inference identifies the causal effect of a variable by answering hypothetical "what if" questions based on observed data [25]. In the factual scenario, the value of Y is formalized as:

$$\begin{aligned} Y_{x,m} &= Y_{x,M_x} = Y(X = x, M = m), \\ m &= M_x = M(X = x). \end{aligned} \tag{1}$$

where X is set to x and M is set to m. In the counterfactual scenario, X is set to different values to assess how this affects both M and Y. As shown in Figure 2b, when X is set to x^*, the mediator M changes, and we obtain $Y_{x^*,m^*} = Y_{x^*,M_{x^*}} = Y(X = x^*, M = M(X = x^*))$. Similarly, Y_{x,m^*} describes the counterfactual scenario where $X = x$ and m^* is set to the value when $X = x^*$.

Causal effects. Causal effects reveal the comparisons between two corresponding outcomes when the value of the reference variable changes [27]. For example, let $X = x$ denote the "taken medication" scenario and $X = x^*$ represent the "not taken medication" scenario. According to causal theory [27], the total effect (TE) of $X = x$ on Y compares two hypothetical situations $X = x$ and $X = x^*$ and is formulated as:

$$TE = Y_{x,M_x} - Y_{x^*,m^*}. \tag{2}$$

The total effect can be divided into two components [28]: the natural direct effect (*NDE*) and the total indirect effect (*TIE*). *NDE* represents the effect of X on Y while the mediator M is held blocked. Therefore, the *NDE* reflects the direct impact of X on Y, excluding any indirect effects mediated by M, with M set to the value it would have when $X = x^*$:

$$NDE = Y_{x,M_{x^*}} - Y_{x^*,M_{x^*}}. \tag{3}$$

TIE is calculated by subtracting the *NDE* from *TE* [28]:

$$TIE = TE - NDE = Y_{x,M_x} - Y_{x,M_{x^*}}. \tag{4}$$

Finally, we obtain TIE as the debiased result by eliminating the direct bias (*NDE*) and use it as the model's final output. Unlike conventional MER methods that derive outputs solely from t, a, and v, CausalMER leverages TIE to effectively eliminate direct biases.

4. Proposed Method: CausalMER

In this section, we first formulate the MER task as a causal graph to describe the causal effects related to modality bias. We then introduce our ***CausalMER*** approach, which eliminates modality bias through counterfactual reasoning.

4.1. Causal-Effect Look at CausalMER

Traditional MER Task. A MER dataset with N samples can be expressed as $\mathcal{D} = \{(\mathcal{T}_1, \mathcal{A}_1, \mathcal{V}_1, \mathcal{Y}_1), \ldots, (\mathcal{T}_N, \mathcal{A}_N, \mathcal{V}_N, \mathcal{Y}_N)\}$. Let $(\mathcal{T}_i, \mathcal{A}_i, \mathcal{V}_i)$ represent a pair of data in a MER dataset, where $\mathcal{T}_i$, $\mathcal{A}_i$, and $\mathcal{V}_i$ correspond to the i-th text, audio, and video data sample, respectively. $\mathcal{Y}_i$ is the emotion labels associated with this i-th data pair. The goal of MER is to train a multimodal model $\mathcal{F}$, which jointly integrates and leverages the three modalities. Therefore, the model's prediction in MER can be represented as $\hat{y}_i = \mathcal{F}(\mathcal{T}_i, \mathcal{A}_i, \mathcal{V}_i)$, where $\hat{y}_i$ is the predicted emotion label for the i-th sample.

Causal Graph in MER. The causal graph of MER is illustrated in Figure 3a. In this graph, there are five variables: text modality T, audio modality A, video modality V, multimodal representation M, and emotional prediction Y, respectively. The effect of T, A, and V on Y can be broken down into two components: the single-modal impact and the multimodal impact. All causal relationships among these variables are explained as follows:

- **Link $T/A/V \rightarrow Y$** represents the direct effect of each modality T, A, and V on the model's prediction Y. This occurs because each modality may contribute salient features with spurious correlations to the prediction. As shown in the paper [12], the importance of a particular modality can vary at the sample level, and no single modality is consistently dominant in every situation. Therefore, it is crucial to avoid making predictions based solely on the direct effect of any one modality. Such direct effects can lead to negative outcomes, as the model relies on spurious correlations between each modality and the labels rather than capturing multimodal relationships.
- **Link $(T, A, V) \rightarrow M \rightarrow Y$** represents the multimodal impact, which captures the indirect effect of T, A, and V on Y through the multimodal knowledge M. The multimodal representation M is obtained by integrating information from the different modalities. In real-world environments, missing modalities can occur, making it difficult to predict which modality will be absent. Therefore, our model performs better by making predictions based on M, trained on all modalities (V, A, and T) together. As a result, the indirect effect provides a beneficial impact on our model.

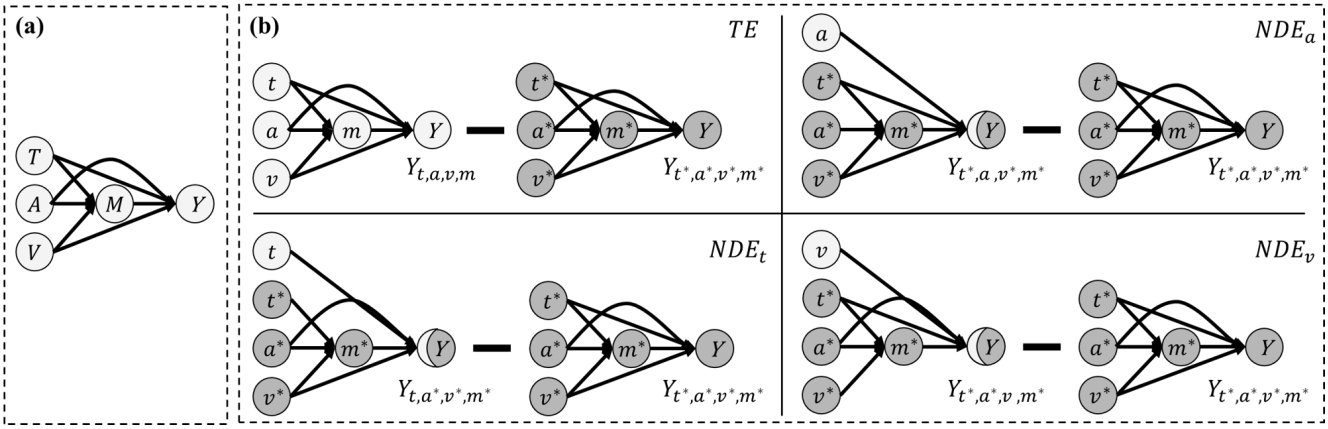

Figure 3. Proposed Method: CausalMER. (**a**) Causal graph for MER. T represents text, A represents audio, V represents video, M denotes multimodal knowledge, and Y is the predicted emotion. (**b**) Counterfactual notations for CausalMER. White nodes represent the values $T = t$, $A = a$, and $V = v$, while gray nodes correspond to the values $T = t^*$, $A = a^*$, and $V = v^*$. Here, * indicates a counterfactual scenario.

In this paper, we propose ***CausalMER*** to reduce modality bias in MER by eliminating the pure language, acoustic, and visual effects that cause certain modalities to dominate the prediction process.

4.2. Counterfactual Inference at CausalMER

Our proposed method aims to mitigate the impact of harmful bias on model predictions when a modality is missing. We represent an emotion Y (e.g., "happy") as the score obtained when T is set to t (e.g., the text "I am so happy"), A is set to a (e.g., a voice saying "I am so happy"), and V is set to v (e.g., a video showing people saying "I am so happy"). Using the counterfactual notations in Equation (1), the causality in the factual scenarios is formulated as follows:

$$Y_{t,a,v,m} = Y(T = t, A = a, V = v, M_{t,a,v} = (T = t, A = a, V = v)). \tag{5}$$

The total effect (*TE*) of *CausalMER* is illustrated in Figure 3. Following the causal effect defined in Equation (2), the TE of $T = t$, $A = a$, and $V = v$ on $Y = y$ is defined as:

$$TE = Y_{t,a,v,m} - Y_{t^*,a^*,v^*,m^*}. \tag{6}$$

where $m^* = M_{t^*,a^*,v^*}$. Here, t^*, a^*, and v^* correspond to the no-treatment condition, representing a counterfactual scenario where the input from specific modalities t, a, and v is blocked. This allows the measurement of the remaining effects, excluding the direct impact. Further details are provided in Section 4.3, Counterfactual Scenario.

In real-world scenarios, it is impossible to predict which modality might be missing. If the model is heavily biased toward a particular modality that is absent, its performance and generalization ability will degrade. Additionally, MER models are prone to spurious correlations between individual modalities and the target emotional labels, preventing the effective use of multimodal knowledge. To address this, we aim to eliminate the direct effect of individual modalities on the emotions. Specifically, we estimate the causal effect of $T = t$, $A = a$, and $V = v$ on $Y = y$ by blocking the impact of M. Figure 3b illustrates the natural direct effects (*NDEs*) of each modality—text, audio, and video—as proposed in our method.

To calculate the direct effect of the text modality T, we apply an intervention [27] that isolates and measures the *NDE* of T on Y. This process involves assessing the impact of the text modality in the direct effect while excluding its contribution to the indirect effect. When $T = t$, $A = a$, and $V = v$, the intervention sets the multimodal representation M to a

specific value m, effectively blocking the mediator M from responding to its inputs. As a result, when the model predicts Y, it relies solely on the text modality T. For the indirect effect, a no-treatment condition t^* is applied. Since the effect of T is blocked for the mediator M with M taking the value $m^* = M_{t^*,a^*,v^*}$, the model explicitly captures the text bias.

$$NDE_t = Y_{t,a^*,v^*,m^*} - Y_{t^*,a^*,v^*,m^*}. \tag{7}$$

Similarly, the NDE of A on Y is calculated as follows:

$$NDE_a = Y_{t^*,a,v^*,m^*} - Y_{t^*,a^*,v^*,m^*}. \tag{8}$$

where NDE_a explicitly captures audio modality bias by isolating the direct effect of A while preventing its impact on the intermediate variable M. The NDE of V on Y is calculated as:

$$NDE_v = Y_{t^*,a^*,v,m^*} - Y_{t^*,a^*,v^*,m^*}. \tag{9}$$

where NDE_v captures video modality bias in a similar manner to NDE_a by isolating the direct effect of V while preventing its influence on M. Finally, the total NDE can be expressed by summing the NDE of each (T, A, V) modality as follows:

$$\begin{aligned} NDE &= NDE_t + NDE_a + NDE_v \\ &= Y_{t,a^*,v^*,m^*} + Y_{t^*,a,v^*,m^*} + Y_{t^*,a^*,v,m^*} - 3 \cdot Y_{t^*,a^*,v^*,m^*}. \end{aligned} \tag{10}$$

The notation for all $NDEs$ is illustrated in Figure 3b. Additionally, modality bias can be mitigated by computing the difference between the TE and NDE, as follows:

$$\begin{aligned} TIE &= TE - NDE \\ &= Y_{t,a,v,m} - Y_{t,a^*,v^*,m^*} - Y_{t^*,a,v^*,m^*} - Y_{t^*,a^*,v,m^*} + 2 \cdot Y_{t^*,a^*,v^*,m^*}. \end{aligned} \tag{11}$$

For inference, the prediction with the maximum TIE is selected as the unbiased result.

4.3. Implementation of CausalMER

As illustrated in Figure 4, we implement CausalMER framework based on the causal graph depicted in Figure 3b.

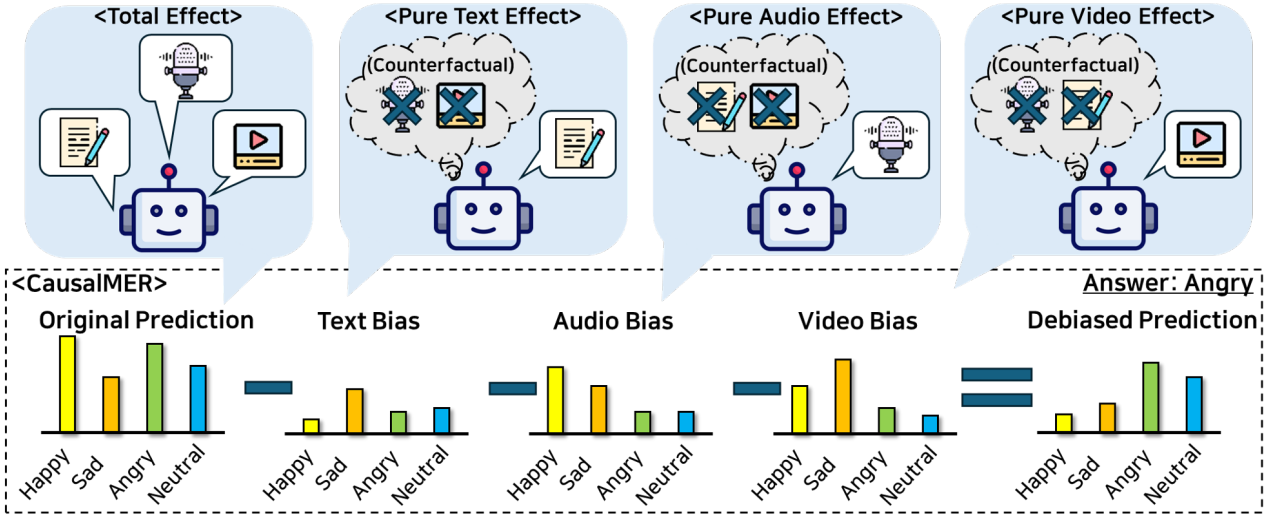

Figure 4. The framework of CausalMER. CausalMER aims to eliminate individual modality bias through counterfactual reasoning. By subtracting the bias from text, audio, and video modalities, CausalMER effectively generates unbiased predictions.

Model Architecture. To determine the effects of text T, audio A, video V, and multimodal M on Y, we utilize four neural models: the text-only branch $\mathcal{F}_T$ (e.g., ALBERT [29]), the audio-only branch $\mathcal{F}_A$ (e.g., ResNet18 [30]), the video-only branch $\mathcal{F}_V$ (e.g., ResNet18 [30]), and multimodal branch $\mathcal{F}_M$. Since the proposed method is model-agnostic, various existing MER approaches can be utilized in the multimodal branch $\mathcal{F}_M$. As detailed in Section 5.3, we use MulT [20], PMR [23], and DMD [24] as the MER backbones for $\mathcal{F}_M$. Our methods are defined as follows:

$$\begin{aligned} Y_t = \mathcal{F}_T(t), Y_a = \mathcal{F}_A(a), Y_v = \mathcal{F}_V(v), Y_m = \mathcal{F}_M(t,a,v), \\ Y_{t,a,v,m} = h(Y_t, Y_a, Y_v, Y_m). \end{aligned} \tag{12}$$

where Y_t, Y_a, Y_v, and Y_m represent the text-only effect ($T \rightarrow Y$), audio-only effect ($A \rightarrow Y$), video-only effect ($V \rightarrow Y$), and the multimodal effect ($T/A/V \rightarrow M \rightarrow Y$), respectively. h is the fusion function to obtain the final score $Y_{t,a,v,m}$.

Counterfactual Scenario. The no-treatment condition is defined as blocking the direct effect of the single modality—text, audio, or video—i.e., when t, a, or v is not provided. These no-treatment conditions are expressed as $T = t^* = \emptyset$, $A = a^* = \emptyset$, and $V = v^* = \emptyset$. However, neural models are unable to handle scenarios where inputs are absent. Therefore, we established the no-treatment condition with the approach used in this study [13]. The terms Y_t, Y_a, Y_v, and Y_m in Equation (12) are expressed as follows:

$$Y_t = \begin{cases} y_t = \mathcal{F}_T(t) & \text{if } T = t \\ y_t^* = c_t & \text{if } T = \emptyset \end{cases},$$

$$Y_a = \begin{cases} y_a = \mathcal{F}_A(a) & \text{if } A = a \\ y_a^* = c_a & \text{if } A = \emptyset \end{cases},$$

$$Y_v = \begin{cases} y_v = \mathcal{F}_V(v) & \text{if } V = vs. \\ y_v^* = c_v & \text{if } V = \emptyset \end{cases},$$

$$Y_m = \begin{cases} y_m = \mathcal{F}_M(t,a,v) & \text{if } T = t \text{ and } A = a \text{ and } V = vs. \\ y_m^* = \text{sum}(c_t, c_a, c_v) & \text{if } T = \emptyset \text{ or } A = \emptyset \text{ or } V = \emptyset \end{cases}. \tag{13}$$

where c_t, c_a, and c_v are learnable parameters.

Fusion Strategies. We combined Y_t, Y_a, Y_v, and Y_m to derive a fused score $Y_{t,a,v,m}$, which can be applied to various fusion strategies. The differences between these strategies are compared in Section 5.5.

- SUM [13] calculates the final score by summing the four probabilities, where σ is the sigmoid function.

$$h(Y_t, Y_a, Y_v, Y_m) = \log \sigma(Y_t + Y_a + Y_v + Y_m). \tag{14}$$

- HM (Harmonic) [13] utilizes the product of sigmoid functions as the fusion method.

$$\begin{aligned} h(Y_t, Y_a, Y_v, Y_m) = \log \frac{Y_{\text{HM}}}{1 + Y_{\text{HM}}}, \\ \text{where } Y_{\text{HM}} = \sigma(Y_t) \cdot \sigma(Y_a) \cdot \sigma(Y_v) \cdot \sigma(Y_m). \end{aligned} \tag{15}$$

- SUM-tanh [9] uses the sum of tanh function, combined with Y_m as the fusion function.

$$h(Y_t, Y_a, Y_v, Y_m) = Y_m + tanh(Y_t) + tanh(Y_a) + tanh(Y_v). \tag{16}$$

- MASK [14] uses the sigmoid function of the sum of Y_t, Y_a, and Y_v as the mask.

$$h(Y_t, Y_a, Y_v, Y_m) = Y_m \cdot \sigma(Y_t + Y_a + Y_v) \tag{17}$$

Training. We apply cross-entropy loss $\mathrm{CE}(\cdot)$ to Y_t, Y_a, Y_c, and $Y_{t,a,v,m}$ to update the model in Equation (12) as follows:

$$\mathcal{L}_{cls} = \mathrm{CE}(Y_{t,a,v,m}, y) + \mathrm{CE}(Y_t, y) + \mathrm{CE}(Y_a, y) + \mathrm{CE}(Y_v, y) \tag{18}$$

where y is the emotion label for (t, a, v).

Following previous work [13], we use Kullback–Leibler divergence $\mathrm{KL}(\cdot)$ to learn the parameters c_t, c_a, and c_v:

$$\mathcal{L}_{kl} = \mathrm{KL}(TE, NDE) \tag{19}$$

where the TE and NDE are calculated by Equation (6) and Equation (10), respectively.

In contrast to prior studies [13], which deal with two modalities, our approach involves three modalities: text, audio, and video. This requires careful balancing of the NDE and TE. The presence of three NDEs introduces challenges in ensuring stable convergence during model optimization. To address this, we introduce an additional loss $\mathcal{L}_{kl_{bal}}$ to balance each NDE and TE, promoting more stable training.

$$\begin{aligned}\mathcal{L}_{kl_{bal}} = {} & \mathrm{KL}(NDE_t, NDE_a) + \mathrm{KL}(NDE_a, NDE_v) + \mathrm{KL}(NDE_v, NDE_t) \\ & + \mathrm{KL}(TE, NDE_t) + \mathrm{KL}(TE, NDE_a) + \mathrm{KL}(TE, NDE_v).\end{aligned} \tag{20}$$

where TE, NDE_t, NDE_a, and NDE_v can be calculated by Equations (6), (7), (8), and (9), respectively. The final loss for CausalMER is expressed as:

$$\mathcal{L} = \sum_{(t,a,v,y)\in\mathcal{D}} \alpha \cdot \mathcal{L}_{cls} + \beta \cdot (\mathcal{L}_{kl} + \mathcal{L}_{kl_{bal}}). \tag{21}$$

where α and β are the hyperparameters.

Inference. According to Equation (11), we utilize TIE as multimodal emotion prediction to eliminate the single-modal impact, as follows:

$$\begin{aligned}TIE &= TE - NDE \\ &= h(Y_t, Y_a, Y_v, Y_m) - h(Y_t, Y_{a^*}, Y_{v^*}, Y_{m^*}) - h(Y_{t^*}, Y_a, Y_{v^*}, Y_{m^*}) \\ &\quad - h(Y_{t^*}, Y_{a^*}, Y_v, Y_{m^*}) + 2 \cdot h(Y_{t^*}, Y_{a^*}, Y_{v^*}, Y_{m^*}).\end{aligned} \tag{22}$$

5. Experiments

In this section, we evaluate CausalMER on two widely recognized benchmark datasets for MER. Our goal is to compare CausalMER with prior competitive approaches and demonstrate its robustness to missing modalities.

5.1. Datasets and Evaluation Metrics

We conducted experiments on two widely used datasets: IEMOCAP [31] and CMU-MOSEI [32]. Refer to Table 1 for a detailed summary of the statistics and evaluation metrics for each dataset.

- **IEMOCAP**: The Interactive Emotional Dyadic Motion Capture (IEMOCAP [31]) dataset contains recordings of 10 actors engaged in conversations, capturing text, audio, and video modalities. It consists of 4453 video clip samples, with pre-split data including 2717 training samples, 798 validation samples, and 938 testing samples, as outlined in previous works [20,23]. The dataset comprises nine emotions (anger, happiness, sadness, neutrality, excitement, frustration, fear, surprise, and others). Following previous works [20,23,33], four emotions (happiness, sadness, anger, and neu-

trality) were chosen for evaluation. Moreover, unlike the CMU-MOSEI [32] dataset, the IEMOCAP [31] dataset adapts a multi-label configuration, allowing video clips to simultaneously exhibit multiple emotions, such as both anger and sadness. As in previous works [20,23], we evaluate the predicted values for each emotional label using binary classification accuracy and F1 score.

- **CMU-MOSEI**: The CMU-MOSEI [32] dataset consists of video clips of movie reviews sourced from YouTube, providing text, audio, and video modality data for each instance. It contains 22,777 video clips, with pre-split data comprising 16,265 training samples, 1869 validation samples, and 938 testing samples, as outlined in previous work [20]. The dataset is assigned sentiment scores ranging from −3 (strongly negative) to +3 (strongly positive). Previous works [20,23,24] have used 7-class accuracy (Acc7) to evaluate model performance.

Table 1. Statistics and evaluation metrics of experimental datasets.

Dataset	Samples			Classes	Metric
	Train	Val	Test		
IEMOCAP [31]	2717	798	938	4	Accuracy (Acc), F1 score (F1)
CMU-MOSEI [32]	16,265	1869	4643	7	

5.2. Implementation Details

Feature Extraction. For the CMU-MOSEI dataset [32], features extracted from the raw data are publicly available due to privacy considerations. The text modality is derived from video transcripts and converted into 300-dimensional vectors using pre-trained GloVe word embeddings (glove.840B.300d) [34]. Facial muscle movements in video frames are captured by extracting 35 facial action units using Facet [35]. Audio features are processed with COVAREP [36], including 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking, and other acoustic parameters, resulting in a 74-dimensional vector. Following previous work [20], we applied the same feature extraction methods used for the CMU-MOSEI dataset to the IEMOCAP dataset.

Experimental Setting. In this study, we conducted experiments using three baseline MER architectures: MulT [20], PMR [23], and DMD [24]. We re-implemented these three baselines based on the publicly available code and integrated them with our CausalMER framework. In all MER backbones, we utilized the pre-trained ALBERT [29] for the text-only branch and ResNet-18 [30] for both the audio-only and video-only branches, respectively. We used the Adam optimizer, and due to the differing characteristics of the IEMOCAP [31] and CMU-MOSEI [32] datasets, we set different learning rates for each. In the IEMOCAP [31] dataset, the text-only branch uses a learning rate of 1×10^{-6}, while the audio-only, video-only, and MER backbone branches use 2×10^{-3}. In the CMU-MOSEI [32] dataset, we trained MulT [20] and PMR [23] with a learning rate of 1×10^{-3}, while DMD [24] was trained with a learning rate of 1×10^{-4}. For the IEMOCAP [31] dataset, we used a batch size of 32. In the CMU-MOSEI [32] dataset, the batch size was 32 for PMR [23] and 16 for both MulT [20] and DMD [24]. In contrast to various counterfactual inferences that use a grid search strategy [10,11,37], CausalMER does not use such an approach. All experiments were implemented using PyTorch (1.13.0) on an NVIDIA RTX A6000 GPU with 48 GB of memory. Additionally, we reported CausalMER with MASK fusion as a key result.

5.3. Main Results

We conducted two main experiments: a standard benchmark in MER and an evaluation of missing modality scenarios on the IEMOCAP [31] and CMU-MOSEI [32] datasets. We aim to assess the efficiency of our CausalMER in a standard benchmark setting and evaluate its robustness in scenarios where a modality is missing. In real-world situations,

the absence of certain modalities can occur, and our CausalMER is designed to address this issue by preventing reliance on any single modality.

CausalMER is evaluated against recent state-of-the-art (SOTA) approaches in MER, with the methods grouped based on the dataset. On the IEMOCAP [31] dataset, we compare CausalMER with EF-LSTM [20], LF-LSTM [20], RMFN [38], MFM [39], RAVEN [33], MCTN [40], MulT [20], PMR [23], FDMER [22], and DMD [24]. For the CMU-MOSEI dataset, we compared our method with DCCA [41] and DCCAE [42], which address missing modalities in a non-recovery manner, and MCTN [40], MMIN [43], GCNet [44], DiCMoR [45], and IMDer [46]. Similar to previous studies [47,48] that have re-run MulT, PMR, and DMD, we reimplemented it following the settings described in the original paper. We report the best performance achieved in our experiments with the configurations provided by the original authors.

Table 2 illustrates the comparison on the IEMOCAP [31] dataset. We integrated the CausalMER framework into MulT [20], one of the most well-recognized models in the MER task. As a result, MulT [20] integrated with CausalMER outperforms the vanilla model in terms of both accuracy and F1 score across all emotions. In particular, the model with CausalMER achieves better results, with improvements of 1.6% in average emotion accuracy and 1.9% in F1 score. This improvement may be attributed to the fact that the original MulT [20] tends to rely on certain single modalities, which reduces the contributions of other less dominant modalities. CausalMER helps address this imbalance. In contrast, CausalMER shows performance improvements, effectively addressing modality bias and balancing modality contributions.

Table 2. Comparison on the IEMOCAP [31] dataset. † indicates our reimplementation.

Model	Happy		Sad		Angry		Neutral		Average	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
EF-LSTM [20]	86.0	84.2	80.2	80.5	85.2	84.5	67.8	67.1	79.8	79.1
LF-LSTM [20]	85.1	86.3	78.9	81.7	84.7	83.0	67.1	67.6	79.0	79.7
RMFN [38]	87.5	85.8	83.8	82.9	85.1	84.6	69.5	69.1	81.5	80.6
MFM [39]	90.2	85.8	88.4	86.1	87.5	86.7	72.1	68.1	84.6	81.7
RAVEN [33]	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3	81.9	81.2
MCTN [40]	84.9	83.1	80.5	79.6	79.7	80.4	62.3	57.0	76.9	75.0
MulT [20]	90.7	88.6	86.7	86.0	87.4	87.0	72.4	70.7	84.3	83.1
MulT † [20]	85.6	82.3	84.0	84.4	88.1	87.8	69.5	69.5	81.8	81.2
+ CausalMER	87.3	86.0	87.3	87.1	88.3	88.2	70.9	71.1	83.4	83.1
(Difference)	+1.7	+3.7	+3.3	+2.7	+0.2	+0.4	+1.4	+1.6	+1.6	+1.9

Table 3 presents the standard benchmarks for the IEMOCAP dataset and the performance drop observed when the available modalities are limited, i.e., when certain modalities are missing. We use MulT [20] as the backbone for CausalMER.

Table 3. The performance drop under six possible missing-modality conditions and the full-modality condition (i.e., "{a}" indicates that only the audio modality is available, with both text and video modalities missing) on the IEMOCAP [31] dataset, utilizing the average accuracy metric. Performance drops that are minimal are highlighted in bold. † indicates our reimplementation.

Model	Performance	Performance Drop Gap							
	{t,a,v}	{a,v}	{t,v}	{t,a}	{t}	{a}	{v}	avg	std
MulT † [20]	81.8	−**2.1**	−6.9	−1.2	−8.8	−**4.1**	−**9.5**	−**5.4**	**3.5**
+ CausalMER	83.4	−5.5	−**4.4**	−1.2	−**6**	−7.8	−11.8	−6.1	**3.5**

When one modality is absent in MulT [20], the drop in performance is more pronounced with the absence of the audio modality than with the absence of the text or video modalities. This is due to MulT [20] being trained dependent on the audio modality. Nonetheless, when integrated with CausalMER, it exhibited a reduced performance drop compared to MulT [20] in the absence of the audio modality, and when the video modality was absent, both demonstrated equivalent performance degradation; however, MulT [20] attained 73% performance, while CausalMER reached 77.4%, thereby achieving superior performance. This demonstrates that CausalMER has effectively eliminated the biases across various modalities. Moreover, although the average performance decline under the missing-modality condition is less pronounced for MulT [20], CausalMER exhibits a higher average performance of 77.3% under the same condition, surpassing MulT's [20] 75.9% by 1.4%. CausalMER outperforms the current MER backbone in scenarios involving modality absence.

Table 4 presents the standard benchmarks for the CMU-MOSEI [32] dataset and the performance drop observed when the available modalities are limited, i.e., when certain modalities are missing. We use MulT [20], PMR [23], and DMD [24] as the backbone for CausalMER.

Table 4. The performance drop under six possible missing-modality conditions and the full-modality condition (i.e., "{a}" indicates that only the audio modality is available, with both text and video modalities missing) on the CMU-MOSEI dataset, utilizing the average accuracy metric. Performance drops that are minimal are highlighted in bold. † indicates our reimplementation.

Model	Performance	Performance Drop Gap						
	{t, a, v}	{a, v}	{t, v}	{t, a}	{t}	{a}	{v}	avg	std
DCCA [41]	47.7	−6.2	−1.1	−1.0	−1.0	−6.6	−6.4	−3.7	2.9
DCCAE [42]	48.2	−6.6	−1.1	−0.8	−1.2	−7.3	−8.1	−4.2	3.5
MCTN [40]	51.2	−9.1	−0.8	−0.5	−1.0	−9.8	−9.6	−5.1	4.8
MMIN [43]	52.4	−10.6	−1.2	−0.4	−1.0	−12	−11.7	−6.2	5.8
GCNet [44]	51.5	−9.5	−0.4	−0.2	−0.3	−10.4	−9.8	−5.1	5.3
DiCMoR [45]	53.4	−11.0	−0.4	−0.7	−1.0	−12.0	−11.4	−6.1	5.9
IMDer [46]	53.4	−10.6	−0.3	−0.3	−0.9	−11.7	−10.8	−5.8	5.8
MulT † [20]	48.8	**−7.4**	**0.3**	**−1.0**	**−1.1**	**−7.4**	**−7.4**	**−4.0**	3.8
+ CausalMER	50.1	−8.0	−1.0	−1.2	−2.8	−8.7	−8.5	−5.0	**3.7**
PMR † [23]	50.3	−11.3	−2.5	−2.0	**−3.6**	−8.9	−10.7	−6.5	4.3
+ CausalMER	48.8	**−6.7**	**0.1**	**−0.3**	−7.2	**−7.4**	**−0.8**	**−3.7**	**3.7**
DMD † [24]	50.8	−9.1	**−1.5**	−2.2	**−3.4**	−9.5	−9.3	−5.8	3.8
+ CausalMER	48.8	**−7.6**	−2.2	**−1.6**	−4.3	**−7.4**	**−7.4**	**−5.1**	**2.8**

Many multimodal models often rely on specific modalities [6], leading to varying performance drops when a modality is missing. For instance, in PMR [23], omitting the video modality resulted in a 2% performance drop, while the absence of the text modality caused an 11.3% decrease—a 9.3% difference. In contrast, when integrating PMR [23] with CausalMER, the video absence led to only a 0.3% drop, and the text absence resulted in a 6.7% reduction—a 6.4% gap, which is 2.9% smaller than the original PMR [23]. This phenomenon occurs because CausalMER calculates the pure modality effect and eliminates modality bias, leading to a more balanced contribution among modalities. As a result, the performance drops are more evenly distributed when certain modalities are absent.

This is evident in the standard deviation of performance across modalities. A smaller standard deviation when each modality is unavailable indicates greater robustness to missing modalities. Integrating MulT [20], PMR [23], and DMD [24] with CausalMER reduces the standard deviation compared to their respective backbones, with standard deviations of 3.7, 3.7, and 2.8, respectively. Furthermore, the integration of the DMD [24] with CausalMER shows the minimal standard deviation compared to all other models in

Table 4. This demonstrates that, despite the backbone of CausalMER not being specifically designed to handle situations where certain modality data are missing, it effectively eliminates modality bias, ensuring balanced contributions from each modality.

5.4. Analysis

We conducted further analysis on CausalMER to evaluate its efficiency in eliminating the single-modal impact. To achieve this, we measured modality contributions using the modality contribution metric [12]. The modality contribution metric, inspired by Shapley theory [49], is proposed to measure the contribution of each modality to the prediction for each sample. A higher modality contribution score indicates a greater contribution of the modality to the prediction, while a lower score suggests a smaller impact. The score is calculated based on the number of correct predictions across all permutations of the modality input.

Table 5 compares the modality contributions of the MER and CausalMER models on the CMU-MOSEI [32] dataset. Each column represents the contribution of a specific modality, with values normalized to sum to 1 across the three modalities. The balance of modality contributions is measured by the standard deviation of these normalized values.

Table 5. Modality contribution in CMU-MOSEI dataset. Normalized Modality Contributions with minimal values are highlighted in bold. † means our reimplementation.

Model	Normalized Modality Contribution			
	Text	**Audio**	**Video**	**std**
MulT † [20]	0.287	0.304	0.409	0.066
+ CausalMER	0.330	0.355	0.315	**0.020**
PMR † [23]	0.393	0.307	0.300	0.051
+ CausalMER	0.323	0.346	0.332	**0.012**
DMD † [24]	0.349	0.316	0.336	**0.017**
+ CausalMER	0.377	0.304	0.319	0.038

When integrating CausalMER with MulT [20] and PMR [23], the standard deviations of modality contributions decrease from 0.066 and 0.051 to 0.02 and 0.012, respectively. This indicates that integrating CausalMER results in more balanced contributions from each modality, effectively eliminating the direct impact of single modality. For DMD [24], the standard deviation is initially low at 0.017. This can be attributed to DMD's [24] use of graph distillation, which evenly distributes knowledge across modalities to balance contributions. Additionally, unlike previous MER methods based on counterfactual inference methods, CausalMER does not employ a grid search algorithm. Grid search is typically used to optimize the weights for each *NDE* in counterfactual inference. While effective for dataset-specific optimization, it is less generalizable to real-world data. As a result, integrating CausalMER with DMD [24] slightly increased the standard deviation by 0.021. Nevertheless, the standard deviation remains lower compared to other backbone models.

5.5. Ablation Studies

We conducted ablation studies on the IEMOCAP [31] dataset, carrying out comprehensive experiments to assess the necessity of each component in CausalMER.

Importance of Components in CausalMER. To assess the effectiveness of our model design, we removed specific components responsible for handling the direct effects of text, audio, and video modalities, as well as the proposed loss function. The results are shown in Table 6. We find that removing the direct effect of each single modality leads to consistent improvement (Rows 1–5).

Table 6. Ablation studies on IEMOCAP dataset. The highest performance is highlighted in bold.

Row	$\mathcal{L}_{kl_{bal}}$	NDE_t	NDE_a	NDE_v	Acc	F1
1	✓	✓	✓	✓	**83.4**	**83.1**
2		✓	✓	✓	82.0	81.9
3	✓	✓	✓		82.3	81.7
4	✓		✓	✓	81.6	81.4
5	✓	✓		✓	80.9	80.7

Unlike previous studies [10,11,13], the proposed CausalMER can identify and remove biases across more than three modalities by using three learnable parameters: c_t, c_a, and c_v. The proposed loss $\mathcal{L}_{kl_{bal}}$ is included to balance the influence of the TE and $NDEs$ during training; otherwise, the model may risk overemphasizing certain terms. Eliminating this loss results in a 1.4% decrease in average accuracy and a 1.2% drop in average F1 score (Row 2), indicating that the loss function effectively promotes balanced learning between the TE and each NDE as intended.

CausalMER effectively balances the effects of the text, audio, and video modalities. To accomplish this, we introduced three modality-specific $NDEs$: NDE_t, NDE_a, and NDE_v. Removing any of these $NDEs$ resulted in performance degradation across all instances (Rows 3–5). This finding suggests that multimodal models are not biased toward a single modality but instead exhibit multiple coexisting modality biases. By successfully mitigating these biases, CausalMER achieved improved performance.

Different Fusion Strategies. Following previous studies [9,13,14], we applied several fusion strategies in Equation (12). Table 7 shows that MASK [14] (Row 1) significantly outperforms other fusion strategies (Rows 2–4), suggesting that this fusion approach effectively mitigates bias across terms in CausalMER. Future research could further explore fusion methods best suited for CausalMER.

Table 7. Impact of different fusion strategies on IEMOCAP [31] dataset.

Row	Fusion Strategy	Acc	F1
1	MASK	83.4	83.1
2	SUM	72.4	74.0
3	HM	71.8	73.1
4	SUM-tanh	74.3	75.5

6. Conclusions and Future Work

Conclusions. CausalMER effectively mitigates modality bias in multimodal emotion recognition, enhancing robustness even when specific modalities are absent. By leveraging causal inference and counterfactual reasoning, it reduces the impact of modality-specific biases without requiring structural changes to existing models. Experimental results demonstrate that CausalMER improves the generalization of MER systems under missing-modality conditions. On the IEMOCAP dataset with the MulT backbone, CausalMER achieves an average accuracy of 83.4%. On the CMU-MOSEI dataset, the average accuracies with MulT, PMR, and DMD backbones are 50.1%, 48.8%, and 48.8%, respectively. These results highlight CausalMER's robustness in missing modality scenarios, as evidenced by its low standard deviation in performance drop gaps. Additionally, our evaluation shows that CausalMER achieves balanced contributions from each modality, effectively mitigating direct biases. CausalMER is particularly suited for real-world applications such as Human–Computer Interaction (HCI), robotics, and healthcare, where real-time emotion recognition is critical. In practical scenarios, sensor failures may lead to missing

modalities, potentially disrupting predictions if the absent modality has a dominant effect. By eliminating modality bias, CausalMER ensures reliable performance, even in such challenging conditions.

Future work. The proposed CausalMER exclusively examines the impact of each modality under full-modality conditions during training. Nevertheless, in different modality-absent scenarios, the modality effect may differ from the full-modality conditions observed. Previous experiments have demonstrated that debiasing derived from the modality effect assessed under full-modality conditions results in strong performance in missing-modality situations. Training the model to account for modality-missing scenarios could enhance its robustness by enabling adaptation to variations in modality effects. Subsequent research will concentrate on broadening this methodology to incorporate these conditions.

Author Contributions: Conceptualization, Y.C. and J.K.; methodology, J.K.; software, J.K.; validation, J.K. and J.H.; formal analysis, J.K. and J.H.; investigation, J.K.; resources, Y.C.; writing—original draft preparation, J.K. and J.H.; writing—review and editing, J.K. and J.H.; visualization, J.K. and J.H.; supervision, Y.C.; project administration, Y.C.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00331, Development of emotion recognition/generation-based interacting edge device technology for mental health care, 50%), by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00437494, 25%) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and by the faculty research fund of Sejong University in 2024 (25%).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data derived from public domain resources. The data presented in this study are available in the public domain. These data were derived from the following resources: IEMOCAP [31] dataset, available at [https://sail.usc.edu/iemocap/ (accessed on 4 December 2024)], and CMU-MOSEI [32] dataset, available at [https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK (accessed on 4 December 2024)]. The versions used in this study were preprocessed and released as part of the MulT [20] study. The specific preprocessed dataset version from MulT [20] can be accessed via [https://github.com/yaohungt/Multimodal-Transformer?tab=readme-ov-file (accessed on 4 December 2024)].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
2. Chen, L.; Li, M.; Wu, M.; Pedrycz, W.; Hirota, K. Coupled Multimodal Emotional Feature Analysis Based on Broad-Deep Fusion Networks in Human–Robot Interaction. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 9663–9673. [CrossRef] [PubMed]
3. Kalateh, S.; Estrada-Jimenez, L.A.; Nikghadam-Hojjati, S.; Barata, J. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access* **2024**, *12*, 103976–104019. [CrossRef]
4. Hudlicka, E. To feel or not to feel: The role of affect in human–computer interaction. *Int. J. Hum. Comput. Stud.* **2003**, *59*, 1–32. [CrossRef]
5. Gibson, K.R.; Gibson, K.R.; Ingold, T. *Tools, Language and Cognition in Human Evolution*; Cambridge University Press: Cambridge, UK, 1993.
6. Hazarika, D.; Li, Y.; Cheng, B.; Zhao, S.; Zimmermann, R.; Poria, S. Analyzing Modality Robustness in Multimodal Sentiment Analysis. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022.
7. Tejero-de Pablos, A. Complementary-Contradictory Feature Regularization Against Multimodal Overfitting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2024.
8. Wu, N.; Jastrzebski, S.; Cho, K.; Geras, K.J. Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022.

9. Chen, Z.; Hu, L.; Li, W.; Shao, Y.; Nie, L. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023.
10. Sun, T.; Wang, W.; Jing, L.; Cui, Y.; Song, X.; Nie, L. Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; Association for Computing Machinery: New York, NY, USA, 2022.
11. Yang, D.; Li, M.; Xiao, D.; Liu, Y.; Yang, K.; Chen, Z.; Wang, Y.; Zhai, P.; Li, K.; Zhang, L. Towards multimodal sentiment analysis debiasing via bias purification. *arXiv* **2024**, arXiv:2403.05023.
12. Wei, Y.; Feng, R.; Wang, Z.; Hu, D. Enhancing Multimodal Cooperation via Sample-level Modality Valuation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024.
13. Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.S.; Wen, J.R. Counterfactual VQA: A Cause-Effect Look at Language Bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021.
14. Zhou, J.; Lin, Y.; Chen, Q.; Zhang, Q.; Huang, X.; He, L. CausalABSC: Causal Inference for Aspect Debiasing in Aspect-Based Sentiment Classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 830–840. [CrossRef]
15. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [CrossRef]
16. Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; Marsic, I. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
18. Dai, W.; Cahyawijaya, S.; Liu, Z.; Fung, P. Multimodal End-to-End Sparse Model for Emotion Recognition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021.
19. Le, H.D.; Lee, G.S.; Kim, S.H.; Kim, S.; Yang, H.J. Multi-Label Multimodal Emotion Recognition with Transformer-Based Fusion and Emotion-Level Representation Learning. *IEEE Access* **2023**, *11*, 14742–14751. [CrossRef]
20. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
21. Liang, T.; Lin, G.; Feng, L.; Zhang, Y.; Lv, F. 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), Online, 11–17 October 2021.
22. Yang, D.; Huang, S.; Kuang, H.; Du, Y.; Zhang, L. Disentangled Representation Learning for Multimodal Emotion Recognition. In Proceedings of the 30th ACM International Conference on Multimedia, New York, NY, USA, 10–14 October 2022.
23. Lv, F.; Chen, X.; Huang, Y.; Duan, L.; Lin, G. 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), Online, 19–25 June 2021.
24. Li, Y.; Wang, Y.; Cui, Z. Decoupled Multimodal Distilling for Emotion Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023.
25. Roese, N.J. Counterfactual thinking. *Psychol. Bull.* **1997**, *121*, 133. [CrossRef] [PubMed]
26. Yang, D.; Yang, K.; Li, M.; Wang, S.; Wang, S.; Zhang, L. Robust Emotion Recognition in Context Debiasing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024.
27. Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* **1986**, *7*, 1393–1512. [CrossRef]
28. Pearl, J. *Causal Inference in Statistics: A Primer*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
29. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations, Online, 26 April–1 May 2020.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vagas, NV, USA, 26 June–1 July 2016.
31. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
32. Bagher Zadeh, A.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018.
33. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.

34. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014.
35. Baltrušaitis, T.; Robinson, P.; Morency, L.P. OpenFace: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016.
36. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014.
37. Tu, G.; Jing, R.; Liang, B.; Yang, M.;Wong, K.F.; Xu, R. A Training-Free Debiasing Framework with Counterfactual Reasoning for Conversational Emotion Detection. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023.
38. Liang, P.P.; Liu, Z.; Bagher Zadeh, A.; Morency, L.P. Multimodal Language Analysis with Recurrent Multistage Fusion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018.
39. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning Factorized Multimodal Representations. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
40. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.P.; Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
41. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep Canonical Correlation Analysis. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013.
42. Wang, W.; Arora, R.; Livescu, K.; Bilmes, J. On Deep Multi-View Representation Learning. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015.
43. Zhao, J.; Li, R.; Jin, Q. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. 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), Online, 1–6 August 2021.
44. Lian, Z.; Chen, L.; Sun, L.; Liu, B.; Tao, J. GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8419–8432. [CrossRef] [PubMed]
45. Wang, Y.; Cui, Z.; Li, Y. Distribution-Consistent Modal Recovering for Incomplete Multimodal Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023.
46. Wang, Y.; Li, Y.; Cui, Z. Incomplete Multimodality-Diffused Emotion Recognition. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.
47. Ma, Y.; Ma, B. Multimodal Sentiment Analysis on Unaligned Sequences Via Holographic Embedding. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 7–13 May 2022.
48. Sahay, S.; Okur, E.; Kumar, S.H.; Nachman, L. Low Rank Fusion based Transformers for Multimodal Sequences. In Proceedings of the Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML), Seattle, WA, USA, 10 July 2020; pp. 29–34.
49. Shapley, L.S. 17. A Value for n-Person Games. In *Contributions to the Theory of Games*; Princeton University Press: Princeton, NJ, USA, 1953; Volume 2.