Causal-Aware Reconstruction Network for Robust Multimodal Emotion Recognition with Missing Modalities

Anonymous ACL submission

Abstract

Multimodal Emotion Recognition (MER) often suffers from missing modalities in real-world scenarios due to sensor malfunctions, asynchronous signals, or degraded inputs. While recent studies have explored modality reconstruction to alleviate this issue, most existing methods rely heavily on dominant co-occurrence patterns in contextual information, which may induce spurious correlations and lead to biased reconstruction results under incomplete modalities. To address this limitation, we introduce a causality-aware perspective into missing-modality emotion recognition. Specifically, we propose a Causal-Aware Reconstruction Network that explicitly models causal cues from conversation history based on the Causal-Cue Encoder to guide the reconstruction process, rather than relying solely on surface-level correlations. Moreover, we design a Granger causality–inspired self-supervised constraint to effectively capture and leverage causal dependencies within multimodal contexts. Extensive experiments on two benchmark datasets demonstrate that our method outperforms existing methods under incomplete modalities.

1 Introduction

Multimodal emotion recognition with missing modalities focuses on real-world scenarios where multimodal inputs are incomplete, such as device malfunctions (Vazquez-Rodriguez et al., 2023; Song et al., 2022), asynchronous signals (Shen et al., 2020; Lin et al., 2023), or low-quality inputs. Given the high prevalence of modality absence in practical multimodal systems, this problem has become a key and rapidly growing research direction in the field of affective computing (Tellamekala et al., 2023; Yuan et al., 2023; Luo et al., 2023).

To address missing-modality challenges, extensive efforts have been made with notable progress (Tran et al., 2017; Cai et al., 2018; Du et al., 2018; Yuan et al., 2021). Most existing methods follow a reconstruction-based paradigm, where available modalities are used to recover missing ones via generative models such as AutoEncoders, VAEs, GANs, and diffusion models (Zhao et al., 2021; Du et al., 2018; Wang et al., 2018, 2024). The reconstructed modalities are then fused to learn joint representations for emotion recognition. To further enhance reconstruction quality, some studies incorporate additional contextual information as auxiliary cues, improving robustness under modality absence (Lian et al., 2023).

Nevertheless, the potential spurious correlations embedded in contextual information can easily interfere with the missing modality reconstruction process and further degrade the performance of downstream emotion recognition. Spurious correlations refer to statistically induced dependencies arising from frequent co-occurrences in contextual information (e.g., repeated utterances, words, or expressions), which exhibit high correlations at the

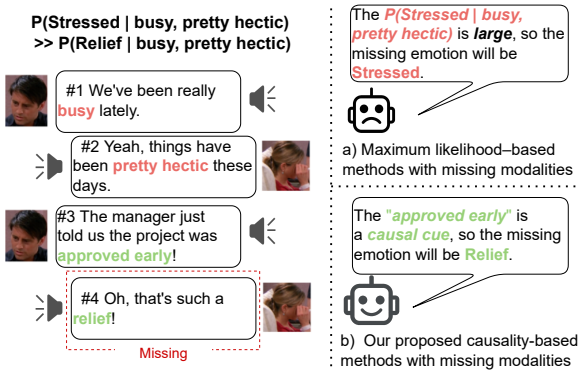


Figure 1: The sample about the spurious correlations. $P(\cdot \mid busy, pretty\ hectic)$ denotes the co-occurrence probability under the given context, where treating high-frequency co-occurrences as causal evidence may induce spurious correlations. a) Conventional maximum likelihood–based methods are interfered with by spurious correlations and therefore reconstruct incorrect information. b) In contrast, our causality-based method is able to correctly reconstruct the missing information.

data level but fail to reflect the true causal mechanisms underlying target representation generation stably (Chen et al., 2024; Yang et al., 2024). As illustrated in Figure 1-(a), under a busy-work background where utterances such as "*busy*" and "*pretty hectic*" frequently co-occur with "*stress*", maximum likelihood–based reconstruction (such as AE, GAN, VAE) tends to favor the dominant response pattern with a high conditional probability. Consequently, when the target emotion is missing, the model is biased toward generating a "*stressed*" response.

Addressing spurious correlations requires uncovering more reliable relationships beyond surface-level co-occurrences in conversations. Motivated by causal inference (Pearl, 2009a, 2010), we argue that causal relationships within conversation history provide more stable guidance for missing-modality reconstruction than purely correlational cues. Based on this insight, we propose a Causal-Aware Reconstruction Network (CARN). Specifically, we first design a *Causal-Cue Feature Encoder* module centered on a causal attention mechanism. By masking future context and sparsifying attention weights, the model is constrained to attend only to historical information that is causally related to the current content. At the same time, the extracted causal cues are explicitly injected into modality-specific representations, thereby suppressing the interference of spurious correlations in both missing-modality reconstruction and emotion prediction. Furthermore, to strengthen the model's capability of perceiving and modeling causal cues, we introduce a *Causal Self-Supervised Loss* based on Granger-causality, which supervises the causal-parent distribution derived from Granger aggregation to align with the causal attention selection. This objective provides effective regularization and training guidance for the *Causal-Cue Feature Encoder* module. In summary, the main contributions of this work are as follows:

- We propose a Causal-Aware Reconstruction Network that explicitly encodes causal cues from historical context and incorporates them into fused representations, therefore mitigating the impact of spurious correlations in conversation history on both modality reconstruction and emotion prediction, and improving model robustness and interpretability.
- We introduce a self-supervised constraint based on Granger causality that simultaneously constrains both causal attention and causal cues, therefore enhancing the model's ability to learn and capture causal dependencies effectively.
- Extensive experiments on benchmark datasets across six missing-modality scenarios demonstrate that our framework substantially outperforms existing models.

2 Related Work

2.1 Causal Inference

Causal inference is a statistical framework that aims to model and reveal the causal dependencies among different factors, thereby identifying the underlying causes of observed outcomes (Pearl, 2009a; Rubin, 2005; Pearl, 2010; Pearl et al., 2016; Yang et al., 2024). Classical frameworks, such as Rubin's causal model (Rubin, 2005) and Pearl's structural causal model (SCM) (Pearl, 2009b), formalize causality through interventions and counterfactual reasoning to estimate causal effects among predefined variables (Yang et al., 2021). As the scope of causal modeling expands to complex data domains such as vision, language, and multimodal learning, observations are often high-dimensional and entangled, making traditional causal models with explicitly predefined variables less directly applicable. This has naturally led to the development of *causal representation learning* (CRL), which aims to recover latent causal factors and their relationships from high-dimensional observations, rather than learning representations that merely reflect statistical correlations (Shen et al., 2022; Yao et al., 2024; Buchholz et al., 2023).

From the perspective of CRL, we propose to capture causal cues from conversation history using a sparse attention matrix, and further integrate these cues into the missing-modality reconstruction process. This design mitigates the interference of spurious correlations in missing-modality reconstruction, while enhancing both the robustness and interpretability of the reconstructed results.

2.2 Self-supervised Learning

Self-supervised learning aims to exploit intrinsic supervisory signals from data itself to learn meaningful representations without relying on explicit annotations (Doersch et al., 2015; Gui et al., 2024). Among various self-supervised paradigms, contrastive learning has emerged as one of the most powerful and widely adopted approaches in recent

years (Liu et al., 2024a; Jia and Liu, 2025; Liu et al., 2025). The central idea is to pull semantically similar samples closer while pushing dissimilar samples apart in the embedding space (Chen et al., 2020; He et al., 2020; Zhang and Stratos, 2021; Pan et al., 2021). This framework was later extended to multimodal settings, where contrastive alignment between different modalities enabled models to capture cross-modal semantics effectively (Wang et al., 2023; Zhou et al., 2024; Song et al., 2024; Dufumier et al., 2025), enhance feature discriminability and improve robustness under noisy or imbalanced conditions (Tu et al., 2023; Hu et al., 2023; Yang et al., 2023).

Despite their effectiveness, existing contrastive learning approaches predominantly focus on correlation-based representation learning. In contrast, our work extends contrastive learning beyond correlational alignment by integrating Granger causality to capture contextual causal cues, thereby improving the interpretability and robustness of missing-modality reconstruction.

3 Methodology

3.1 Architecture Overview

As illustrated in Figure 2, our CARN framework consists of four main components: 1) **Causal-Cue based Feature Encoder (CCE)**, 2) **Feature Fusion**, 3) **Reconstruction Decoder**, and 4) **Classifier**. The **Causal-Cue based Feature Encoder** (Figure 2-a) is first explicitly models intra-modality causal dependencies from historical context by leveraging sparse causal attention and further integrates into modality-specific representations to suppress spurious correlations during feature encoding. Afterthat, the causally enhanced modality features are fed into the **Feature Fusion** module (Figure 2-b) to generate a joint representation, which is simultaneously utilized by a **Classifier** for emotion prediction and a **Reconstruction Decoder** for recovering missing modality features. To further regularize causal cue learning, we introduce a **Causal Self-Supervised Loss** based on Granger causality (Figure 2-c), which encourages the model to identify compact and meaningful causal parents across modalities.

3.2 Causal-aware Reconstruction

3.2.1 Causal-Cue based Feature Encoder

The Causal-Cue based Feature Encoder module is the core component of CARN. It introduces a causal attention mechanism to model causal cues within each modality based on historical information. Notably, to encode modality-specific causal information, we design an independent causal feature encoding channel for each modality.

As illustrated in Figure 2-a, given an arbitrary modality input $x_m \in \mathbb{R}^{T\times d_m}$, where $m \in \{a, t, v\}$, we first compute its causal attention matrix. Specifically, we start by calculating the attention scores for x_m:

$$S_m = \frac{Q_m K_m^\top}{\sqrt{d_m}} \in \mathbb{R}^{T\times T}, \tag{1}$$

where $Q_m = x_m W_Q$, $K_m = x_m W_K$, and W_Q, W_K denote the query and key projection matrices, respectively. In contrast to vanilla attention, we define a causal mask $M \in 0, 1^{T\times T}$ and apply it to the attention scores to restrict the attention scope. This masking operation encourages the model to focus on causal relationships within historical information:

$$M_m^{ij} = \begin{cases} 1, & \text{if } i \geq j, \\ 0, & \text{if } i < j. \end{cases} \tag{2}$$

$$\bar{S}_m^{ij} = \begin{cases} S_m^{ij}, & \text{if } M_m^{ij} = 1, \\ -\infty, & \text{if } M_m^{ij} = 0. \end{cases} \tag{3}$$

where $M_m^{ij} = 1$ indicates that time step i is allowed to attend to time step j, while $M_m^{ij} = 0$ prevents any leakage of future information. $\bar{S}_m^{ij}$ denote the masked attention scores. Furthermore, we apply Entmax-1.5 (Correia et al., 2019) to sparsify the masked attention scores, encouraging the model to focus on capturing potential causal dependencies within historical information while suppressing irrelevant time steps:

$$P_m^{(i,j)} = \bar{S}_m^{(i,j)} - \max_j \bar{S}_m^{(i,j)}, \tag{4}$$

$$A_m^{(i,j)} = \frac{\left[\max(P_m^{(i,j)} - \tau_m, 0)\right]^2}{\sum_j \left[\max(P_m^{(i,j)} - \tau_m, 0)\right]^2}, \tag{5}$$

where τ_m is the threshold parameter of Entmax and A_m denotes the sparsified causal attention matrix. Based on the resulting causal attention distribution A_m, we further compute the modality-specific causal cue z_m:

$$z_m = A_m V_m, \tag{6}$$

where $V_m = x_m W_V$, and W_V denotes the value projection matrix.

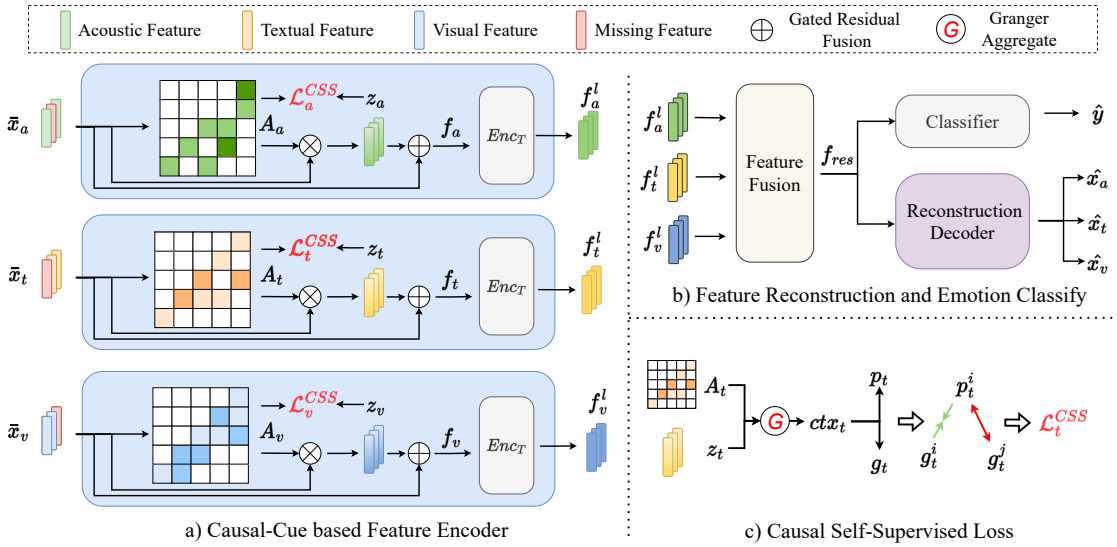


Figure 2: The overview of the Causal-Aware Reconstruction Network. a) The Causal-Cue-based Feature Encoder, which captures causally informative representations within each modality under a causal attention mechanism. b) The feature fusion, reconstruction, and emotion classifier modules, where multimodal causal features are integrated to reconstruct missing modality representations and support emotion prediction. c) The computation pipeline of the Causal Self-Supervised Loss

To better integrate causal cues with the original features, we introduce an adaptive gated fusion mechanism to fuse the original input x_m with the causal cues z_m. Specifically, we first compute a gating coefficient γ_m as:

$$\gamma_m = \sigma\big(\mathbf{W}_m^{\gamma}[x_m; z_m] + b_m^{\gamma}\big), \tag{7}$$

where $[\cdot;\cdot]$ denotes vector concatenation and $\sigma(\cdot)$ is the sigmoid activation function. We then adopt a gated residual fusion scheme to inject the causal features into x_m:

$$f_m = \mathrm{LN}\big(x_m + \gamma_m \odot z_m\big), \tag{8}$$

where $\odot$ denotes element-wise multiplication and $\mathrm{LN}(\cdot)$ denotes Layer Normalization. Subsequently, f_m is fed into a Transformer encoder Enc_T to obtain the latent representation:

$$f_m^l = Enc_T(f_m). \tag{9}$$

3.2.2 Feature Fusion and Classifier

Following prior work (Lian et al., 2023), we concatenate the latent representations from different modalities to form a joint embedding:

$$f_{\text{joint}} = [f_a^l; f_t^l; f_v^l]. \tag{10}$$

To enhance inter-modal interactions and improve information fusion, we further apply a nonlinear transformation to f_{joint} and introduce a residual connection between the transformed features and the original joint representation:

$$f_{\text{res}} = f_{\text{joint}} + \mathrm{ReLU}(\mathrm{MLP}(f_{\text{joint}})). \tag{11}$$

Finally, the fused representation is fed into a classification head to obtain the emotion prediction:

$$\hat{y} = \mathcal{E}^{\text{cls}}(f_{\text{res}}), \tag{12}$$

where $\mathcal{E}^{\text{cls}}(\cdot)$ denotes the emotion classifier.

3.2.3 Feature Reconstruction

The Feature Reconstruction module consists of a Transformer encoder followed by an MLP layer. Specifically, the fused representation f_{res} is first fed into the Transformer encoder, and the resulting output is then mapped by an MLP to obtain:

$$\hat{x} = \mathrm{MLP}\big(\mathrm{TransEnc}(f_{\text{res}})\big) \in \mathbb{R}^{d_a+d_t+d_t}, \tag{13}$$

where $d_a = 512$, $d_t = 1024$, and $d_v = 1024$ denote the feature dimensions of the acoustic, textual, and visual modalities, respectively. Subsequently, $\hat{x}$ is split according to these modality-specific dimensions to obtain the reconstructed features for each modality:

$$\hat{x}_a, \hat{x}_t, \hat{x}_v = \mathrm{Split}(\hat{x}). \tag{14}$$

3.3 Causal Self-Supervised Loss

To further regularize the learning of causal cues, we design a **Causal Self-Supervised** loss inspired

by Granger-style causality (Shojaie and Fox, 2022). The core idea is to encourage each temporal representation within a modality to be predictable from its valid causal parents, while being distinguishable from non-parent contexts.

As shown in Figure 2-c, taking the textual feature sequence $x_t \in \mathbb{R}^{T\times d_t}$ and the corresponding causal attention matrix $A_t \in \mathbb{R}^{T\times T}$ as an example, we first compute a context representation for each time step by aggregating its causal parents:

$$\mathrm{ctx}_t[b,j,:] = \sum_i A_t[b,i,j]\cdot x_t[b,i,:], \quad (15)$$

where $A_t[b,i,j]$ denotes the attention weight from time step i (source) to time step j (target) in batch b. We then apply two projection networks, $\mathcal{F}_t(\cdot)$ and $\mathcal{G}_t(\cdot)$, to map the context representation and the target embedding into a shared latent space:

$$p_t = \mathcal{F}_t(\mathrm{ctx}_t), \quad g_t = \mathcal{G}_t(x_t), \quad (16)$$

where p_t is used to predict the current textual representation from its causal parents, and g_t serves as the ground-truth embedding of the current time step. For training stability, the target embeddings g_t are detached from the gradient flow. Finally, we flatten all valid temporal positions across the batch into N samples and construct an InfoNCE-style contrastive objective:

$$\mathcal{L}_t^{\mathrm{CSS}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left((p_t^i)^\top g_t^i/\tau\right)}{\sum_{j=1}^{N}\exp\left((p_t^i)^\top g_t^j/\tau\right)}, \quad (17)$$

where τ is a temperature hyperparameter controlling the sharpness of the similarity distribution. In this formulation, each prediction derived from causal parents is encouraged to align with its corresponding ground-truth embedding (i.e., the positive pair), while embeddings from other time steps act as negative samples. Minimizing this objective enforces accurate prediction from causally aggregated context, thereby guiding the model to extract meaningful and stable causal cues.

3.4 Training and Joint Optimization

To ensure that each pipeline in CARN learns robust causal cues and emotion information, and can effectively recognize emotions under missing-modality conditions, we adopt a two-stage training strategy inspired by prior work (Zhao et al., 2021; Liu et al., 2024b; Xu et al., 2024).

Stage 1: Training with full modalities. In this stage, full-modality inputs (x_a, x_t, x_v) are fed into the causal-aware reconstruction network, enabling each pipeline to learn complete causal cues (z_a, z_t, z_v) and latent representations (f_a^l, f_t^l, f_v^l). To supervise each modality in capturing emotion-discriminative information, we introduce independent emotion classification heads for each modality and predict emotions using modality-specific features. In addition, to ensure the quality of causal cue learning, we incorporate causal self-supervised losses into the CCE module of each pipeline, constraining the encoder to learn accurate causal relationships. The training objective of this stage jointly optimizes the emotion classification loss $\mathcal{L}_{\mathrm{EMO}}^1$ and the causal self-supervised loss $\mathcal{L}_{\mathrm{CSS}}$:

$$\mathcal{L}_{\mathrm{total}}^1 = \mathcal{L}_{\mathrm{EMO}}^1 + \mathcal{L}_{\mathrm{CSS}}, \quad (18)$$

$$\mathcal{L}_{\mathrm{EMO}}^1 = \sum_{m\in\{a,t,v\}} \mathrm{CE}\big(y, \mathcal{E}_m^{\mathrm{cls}}(f_m^l)\big), \quad (19)$$

$$\mathcal{L}_{\mathrm{CSS}} = \sum_{m\in\{a,t,v\}} \mathcal{L}_m^{\mathrm{CSS}}(x_m, A_m), \quad (20)$$

where $\mathcal{E}_m^{\mathrm{cls}}$ denotes the emotion classification head for modality m, y is the ground-truth emotion label, and $\mathrm{CE}(\cdot)$ denotes the cross-entropy loss.

Stage 2: Training with missing modalities. After Stage 1, the model has learned emotion-discriminative representations and causal cues across modalities. We then fine-tune the pretrained model on modality-missing data to adapt it to incomplete-modality scenarios. Accordingly, the training objective in this stage consists of the emotion recognition loss under missing-modality conditions and the reconstruction loss for missing modalities:

$$\mathcal{L}_{\mathrm{total}}^2 = \mathcal{L}_{\mathrm{EMO}}^2 + \mathcal{L}_{\mathrm{REC}}, \quad (21)$$

$$\mathcal{L}_{\mathrm{EMO}}^2 = \mathrm{CE}(y, \hat{y}), \quad (22)$$

$$\mathcal{L}_{\mathrm{REC}} = \sum_{m\in\{a,t,v\}} \mathrm{MSE}(x_m, \hat{x}_m), \quad (23)$$

where $\mathcal{L}_{\mathrm{REC}}$ is the reconstruction loss based on the mean squared error (MSE).

4 Experiments

4.1 Datasets and Metrics

To validate the effectiveness of our approach, we conducted extensive experiments on two public benchmark datasets:

Dataset	Model	Available Modality													
		a		v		t		av		at		vt		Average	
		WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
IEMOCAP four-class	CRA	0.3756	0.2623	0.3801	0.2607	0.4732	0.3698	0.3758	0.2619	0.4817	0.3806	0.4755	0.3715	0.4270	0.3178
	CPMNet	0.4685	0.5172	0.4495	0.4449	0.4563	0.4532	0.4867	0.4933	0.3481	0.3623	0.4562	0.4657	0.4442	0.4561
	MMIN	0.5658	0.5900	0.5252	0.5060	0.6657	0.6802	0.6399	0.6343	0.7294	0.7114	0.7199	0.6843	0.6410	0.6344
	GCNet	0.6095	0.6504	0.5652	0.5243	0.7421	0.6613	0.6455	0.6515	0.7721	0.7651	0.7505	0.7307	0.6808	0.6639
	CIF-MMIN	0.5753	0.6006	0.5346	0.5156	0.6722	0.6899	0.6499	0.6353	0.7419	0.7259	0.7240	0.6991	0.6497	0.6444
	MoMKE	0.6919	0.7043	0.5641	0.5226	0.7758	0.7790	0.6780	0.6490	0.7906	0.7979	0.7561	0.7459	0.7094	0.6998
	CARN (ours)	**0.7089**	**0.7142**	**0.5812**	**0.5323**	**0.7821**	**0.7856**	**0.6923**	**0.6747**	**0.8051**	**0.8125**	**0.7597**	**0.7507**	**0.7216**	**0.7117**
	δ_{SOTA}	↑0.0170	↑0.0099	↑0.0160	↑0.0080	↑0.0063	↑0.0066	↑0.0143	↑0.0232	↑0.0145	↑0.0146	↑0.0036	↑0.0048	↑0.0121*	↑0.0119*
IEMOCAP six-class	CRA	0.3142	0.4321	0.3121	0.1667	0.3357	0.2830	0.3146	0.1712	0.3367	0.2838	0.3398	0.2856	0.3255	0.2269
	CPMNet	0.2947	0.2980	0.2620	0.2495	0.3244	0.3495	0.2692	0.2546	0.3349	0.3394	0.3134	0.3043	0.2998	0.2992
	MMIN	0.4644	0.4302	0.3463	0.2962	0.4105	0.3536	0.4184	0.4056	0.5133	0.4743	0.4583	0.4021	0.4352	0.3937
	GCNet	0.4911	0.4648	**0.3971**	0.3425	0.5676	0.5770	0.4551	0.4315	0.5834	0.5709	0.5743	0.5403	0.5114	0.4878
	CIF-MMIN	0.4518	0.4438	0.3715	0.3248	0.4345	0.3886	0.4284	0.4293	0.5220	0.4873	0.4678	0.4233	0.4460	0.4162
	MoMKE	0.5039	0.4746	0.3896	0.3388	0.6032	0.5952	0.4807	0.4499	0.6315	0.6193	**0.5959**	0.5659	0.5341	0.5073
	CARN (ours)	**0.5199**	**0.4887**	0.3807	**0.3490**	**0.6073**	**0.6119**	**0.4928**	**0.4648**	**0.6360**	**0.6274**	0.5916	**0.5681**	**0.5381**	**0.5183**
	δ_{SOTA}	↑0.0160	↑0.0141	↓-0.0164	↑ 0.0065	↑ 0.0041	↑0.0167	↑ 0.0121	↑0.0149	↑0.0045	↑0.0081	↓-0.0043	↑0.0022	↑0.0039	↑0.0110 *
Dataset	Model	a		v		t		av		at		vt		Average	
		WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1
CMUMOSEI	CRA	0.5843	0.4321	0.5843	0.4310	0.7177	0.7108	0.5830	0.4310	0.7097	0.7042	0.7184	0.7101	0.6496	0.5699
	CPMNet	0.6571	0.6518	0.6123	0.6173	0.7287	0.7244	0.6156	0.6199	0.7265	0.7224	0.6629	0.6684	0.6672	0.6674
	MMIN	0.6429	0.5897	0.5878	0.5259	0.8272	0.8271	0.6341	0.5901	0.8299	0.8398	0.8470	0.8465	0.7282	0.7032
	GCNet	0.6092	0.6040	0.6899	0.6654	0.8471	0.8443	0.7149	0.7067	0.8538	0.8524	0.8600	0.8596	0.7625	0.7554
	CIF-MMIN	0.6732	0.6471	0.6408	0.6010	0.8376	0.8366	0.6471	0.6291	0.8486	0.8474	0.8536	0.8530	0.7502	0.7357
	MoMKE	**0.7256**	0.7103	0.6632	0.6581	0.8646	0.8643	**0.7237**	0.7207	0.8632	0.8629	0.8690	0.8691	0.7849	0.7809
	CARN (ours)	0.7226	**0.7216**	**0.7014**	**0.6991**	**0.8665**	**0.8666**	0.7232	**0.7239**	**0.8657**	**0.8638**	**0.8718**	**0.8708**	**0.7919**	**0.7910**
	δ_{SOTA}	↓-0.0030	↑0.0113	↑0.0115	↑0.0337	↑0.0019	↑0.0023	↓-0.0005	↑0.0032	↑0.0025	↑0.0009	↑0.0028	↑0.0017	↑0.0070	↑0.0101*

Table 1: The performance comparison with state-of-the-art (SOTA) methods under six fixed missing-modality conditions on two benchmark datasets. "Average“ denotes the mean performance of each model across all six conditions. The best results on each dataset are highlighted in **bold**, while the second-best results are underlined. The row labeled δ_{SOTA} reports the performance gain or loss of our model relative to the strongest baseline. $*$ indicates statistical significance with $p < 0.05$.

IEMOCAP (Busso et al., 2008) is a widely used multimodal emotion recognition dataset. It has often been used in previous works for four-class classification tasks (*Happy, Sad, Neutral, Angry*) (Zhao et al., 2021; Liu et al., 2024b) and six-class classification tasks (*Happy, Angry, Sad, Neutral, Surprised, Fearful*) (Mai et al., 2020; Lian et al., 2023). In this work, we incorporate both of them to evaluate our method against the baseline.

CMU-MOSEI is a large-scale benchmark dataset for multimodal sentiment analysis, comprising 22,856 video clips collected from YouTube, which are split into training, validation, and test sets containing 16,326, 1,871, and 4,659 samples, respectively. Following prior work (Xu et al., 2024), we frame the task as a binary sentiment classification problem, where utterances with sentiment polarity > 0 are labeled as *positive*, and those with polarity < 0 are labeled as *negative*.

For the IEMOCAP dataset, we follow previous work (Xu et al., 2024) and use weighted accuracy (WA) and unweighted accuracy (UA) as evaluation metrics. For the CMU-MOSEI dataset, we use accuracy (Acc) and F1 score as evaluation metrics (Liu et al., 2024b).

4.2 Implementation Details

To evaluate the performance of our model and ensure a fair comparison, we perform experiments using involved six missing modality combinations (Lian et al., 2023; Xu et al., 2024): {a}, {t}, {v}, {a, t}, {a, v}, and {t, v}, where ‘a’, ‘t’, and ‘v’ represent the acoustic, textual, and visual modalities, respectively. Each set indicates the modalities available during inference. For instance, {a} means that only the acoustic modality is available. To ensure fair comparison, we adopt publicly available features from (Lian et al., 2023; Xu et al., 2024). All models were trained for 200 epochs using the Adam optimizer with a learning rate of 0.0001 and a dropout rate of 0.5. To ensure the objectivity of the experimental results, we conducted all experiments three times and reported the average results.

4.3 Overall Comparisons

Table. 1 reports the performance comparison between our method and state-of-the-art (SOTA) baselines, including **CRA** (Tran et al., 2017), **CPMNet** (Zhang et al., 2020), **MMIN** (Zhao et al., 2021),

Dataset	Model	Available Modality													
		a		v		t		av		at		vt		Average	
		WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
IEMOCAP four-class	CARN (ours)	**0.7089**	**0.7142**	**0.5812**	**0.5323**	**0.7821**	**0.7856**	**0.6923**	**0.6747**	**0.8051**	**0.8125**	**0.7597**	**0.7507**	**0.7216**	**0.7117**
	w/o CCE	0.6944	0.7059	0.5722	0.5251	0.7670	0.7758	0.6679	0.6442	0.7826	0.7888	0.7524	0.7428	0.7061	0.6971
	w/o CSS	0.6988	0.7078	0.5672	0.5198	0.7603	0.7685	0.6835	0.6653	0.7896	0.7959	0.7292	0.7128	0.7048	0.6950
	w/ MSE	0.6711	0.6674	0.5641	0.5468	0.7191	0.7193	0.6656	0.6567	0.7568	0.7573	0.7345	0.7346	0.6852	0.6803
	w/ vanilla attention	0.6918	0.6875	0.5474	0.5213	0.7726	0.7714	0.6812	0.6749	0.8017	0.8006	0.7075	0.7007	0.7004	0.6927
IEMOCAP six-class	CARN (ours)	**0.5199**	0.4887	**0.3807**	**0.3490**	**0.6073**	**0.6119**	**0.4928**	**0.4648**	**0.6360**	**0.6274**	**0.5916**	**0.5681**	**0.5381**	**0.5183**
	w/o CCE	0.5164	0.4789	0.3654	0.3459	0.5918	0.5843	0.4630	0.4440	0.6264	0.6171	0.5687	0.5524	0.5220	0.5038
	w/o CSS	0.4905	0.4785	0.3698	0.3347	0.5930	0.5870	0.4838	0.4585	0.6130	0.5994	0.5511	0.5304	0.5169	0.4981
	w/ MSE	0.4918	0.4769	0.3682	0.3161	0.5619	0.5608	0.4712	0.4565	0.5966	0.5945	0.5278	0.517	0.5029	0.487
	w/ vanilla attention	0.5158	**0.4915**	0.3654	0.3223	0.5969	0.5898	0.4904	0.4632	0.6325	0.6274	0.5242	0.5148	0.5236	0.5047
Dataset	Model	a		v		t		av		at		vt		Average	
		WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1
CMUMOSEI	CARN (ours)	**0.7226**	**0.7216**	**0.7014**	**0.6991**	**0.8665**	**0.8666**	0.7232	**0.7239**	**0.8657**	0.8638	**0.8718**	**0.8708**	**0.7919**	**0.7910**
	w/o CCE	0.7200	0.7151	0.6855	0.6825	0.8613	0.8604	**0.7289**	0.7159	0.8643	0.8639	0.8404	0.8385	0.7834	0.7794
	w/o CSS	0.7165	0.7128	0.6932	0.6904	0.8598	0.8590	0.7177	0.7157	0.8652	**0.8646**	0.8690	0.8689	0.7869	0.7852
	w/ MSE	0.6907	0.6891	0.6802	0.6708	0.8407	0.8414	0.6783	0.6786	0.8437	0.8431	0.8423	0.8425	0.7626	0.7609
	w/ vanilla attention	0.6775	0.6759	0.6723	0.6649	0.844	0.8439	0.6868	0.6795	0.8487	0.8476	0.8473	0.8468	0.7628	0.7598

Table 2: The performance of the ablation experiments under six missing conditions.

GCNet (Lian et al., 2023), **CIF-MMIN** (Liu et al., 2024b), and **MoMKE** (Xu et al., 2024), under six fixed missing-modality conditions. Overall, our method achieves the best average performance on all three benchmarks. Across experiments on both datasets, our method outperforms MoMKE by approximately 1% on average, and most of the improvements are statistically significant ($p < 0.05$). It is worth noting that our method achieves significant improvements over GCNet. GCNet leverages two graph-based modules to capture contextual correlations in conversations. In contrast, our proposed causal cue encoder emphasizes extracting causal cues from the context. The superior performance of our method suggests that causal relations are not merely equivalent to contextual correlations.

Microscopically, we analyzed the average accuracy of each emotion category for different models on the IEMOCAP dataset to provide a more in-depth evaluation. The experimental results are shown in Figure 3. From the trend of the Figure 3-a, our method consistently outperforms the baseline methods across the four emotion categories in IEMOCAP four-class, with particularly strong performance in *Sadness*, *Neutral*, and *Happiness*. From Figure 3-b, it can be observed that our method consistently outperforms the baselines across most categories, with the largest improvements seen in *Sadness*. In general, these trends indicate that our method effectively enhances the discriminative ability of the model and maintains stable performance in all categories of emotions.

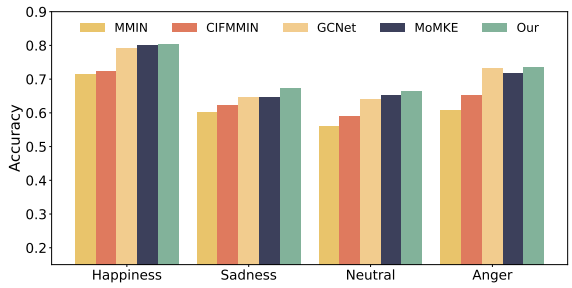
4.4 Ablation Study

To further examine the contribution of each component in CARN, we conduct ablation experiments on the benchmark datasets under six missing-modality conditions.

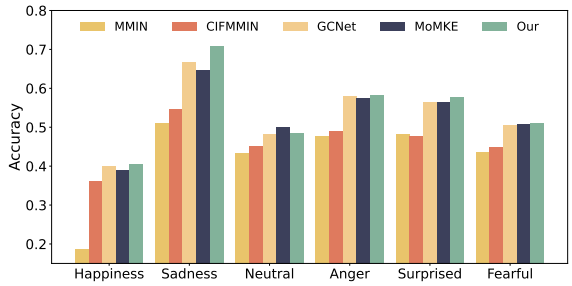
1) w/o Causal-Cue Feature Encoder (w/o CCE): To evaluate the contribution of the causal-cue feature encoder, we remove the CCE module and retain only the Transformer encoder for modality-specific feature extraction. As shown in Table 2, removing CCE results in consistent performance degradation across all missing-modality settings, indicating that explicitly modeling intra-modality causal cues is essential for suppressing spurious correlations and improving robustness under modality absence.

2) w/o Causal Self-Supervised Loss (w/o CSS): To analyze the role of the causal self-supervised objective, we remove the CSS term while keeping other components unchanged. The results in Table 2 show that the removal of CSS leads to noticeable and consistent performance drops across different modality combinations, demonstrating that the self-supervised loss provides stable regularization for causal cue learning and contributes to higher-quality representations

3) w/ MSE Loss: We replace the proposed contrastive self-supervised loss with an MSE loss to study their effects on causal representation learning. As shown in Table 2, this replacement leads to significant performance degradation. This is because MSE optimizes toward conditional expectation, en-



(a) IEMOCAP four-class



(b) IEMOCAP six-class

Figure 3: Performance comparison of different models across emotion categories on the IEMOCAP dataset.

couraging the model to fit average trends rather than learning causally stable representations. In contrast, the proposed contrastive objective maximizes mutual information between causal predictions and corresponding targets while suppressing non-causal associations, which better aligns with causal invariance and independent causal mechanisms. Therefore, contrastive self-supervision is more suitable for constraining causal cue learning.

4) w/ vanilla attention: In Sec. 3.2, we utilize a sparse attention matrix with masked contexts to capture causal relationships within the same modality. To validate the effectiveness of our approach, we replace this attention mechanism with vanilla attention, which incorporates full contextual information. As shown in Table 2, this replacement leads to a significant performance degradation, demonstrating the necessity and effectiveness of the proposed sparse causal attention mechanism.

To further analyze this effect, we visualize the attention matrices of vanilla attention and causal attention in Fig. 4. Vanilla attention exhibits dense and unstructured patterns with unrestricted temporal access, making it prone to spurious correlations and temporal leakage. In contrast, causal attention yields a strictly lower-triangular and sparse structure, attending only to historical timesteps and identifying compact causal parent positions. These visualizations demonstrate that causal attention effectively captures meaningful causal dependencies while improving interpretability and robustness.

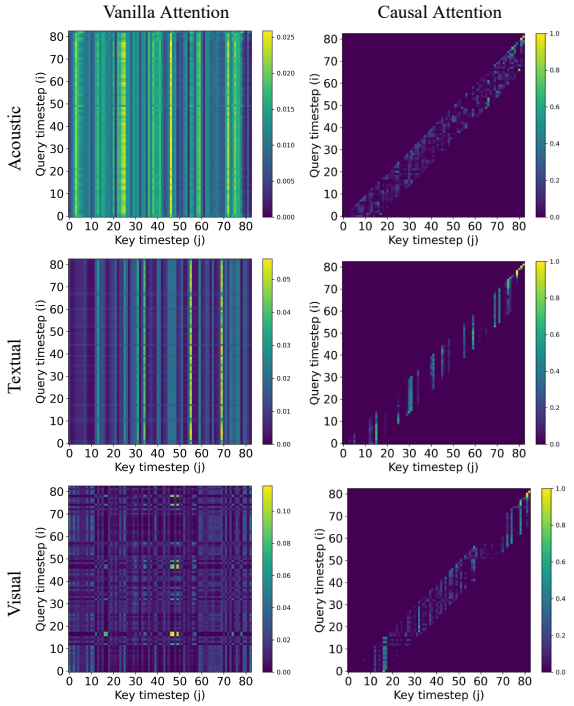


Figure 4: Visualization of attention matrices for various modality pipelines within the CCE module on the IEMOCAP four-class task. The left column shows the vanilla attention, while the right column shows the proposed causal attention.

5 Conclusion

To address the interference caused by spurious correlations in conversational history during missing-modality reconstruction, we propose the Causal-Aware Reconstruction Network (CARN). This framework explicitly models causal cues from conversation history and integrates them into the reconstruction process, enabling the model to rely on more stable and meaningful contextual information rather than superficial correlations. As a result, the robustness and interpretability of reconstructed modalities are significantly enhanced, leading to improved emotion recognition performance under missing-modality conditions. Moreover, we introduce a Granger-causality-based self-supervised constraint to further strengthen the model's ability to perceive and exploit causal cues. Extensive experimental results demonstrate the effectiveness and robustness of the proposed approach across various missing-modality settings.

Limitations

Despite its effectiveness, our approach relies on several assumptions that merit further investigation. First, the causal cues leveraged in this work are derived from Granger causality, which characterizes temporally predictive relationships rather than definitive interventional causal effects. While this formulation is well suited for conversational and sequential data, it may not fully capture more complex causal structures involving latent confounders or instantaneous interactions. Second, our framework primarily exploits historical contextual information for missing-modality reconstruction, and its effectiveness may depend on the availability and quality of such context. In summary, future work may explore more general causal formulations beyond Granger-style temporal dependencies and extend the proposed framework to better accommodate scenarios with limited, noisy, or unreliable contextual information, thereby further enhancing its applicability in real-world settings.

References

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. 2023. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36:45419–45462.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166.

Jiali Chen, Yi Cai, Ruohang Xu, Jiexin Wang, Jiayuan Xie, and Qing Li. 2024. Deconfounded emotion guidance sticker selection with causal inference. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3084–3093.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. 2019. Adaptively sparse transformers. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 2174–2184. Association for Computational Linguistics.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430.

Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 108–116.

Baptiste Dufumier and 1 others. 2025. Comm: Contrastive multimodal learning beyond redundancy. In *Proceedings of the 2025 International Conference on Learning Representations (ICLR)*.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. 2024. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735. IEEE.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhenqi Jia and Rui Liu. 2025. Intra-and inter-modal context interaction modeling for conversational speech synthesis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432.

Wei-Cheng Lin, Lucas Goncalves, and Carlos Busso. 2023. Enhancing resilience to missing data in audio-text emotion recognition with multi-scale chunk regularization. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 207–215.

Rui Liu, Shuwei He, Yifan Hu, and Haizhou Li. 2025. Multi-modal and multi-scale spatial environment understanding for immersive visual text-to-speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24632–24640.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. 2024a. Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18698–18706.

Rui Liu, Haolin Zuo, Zheng Lian, Bjorn W Schuller, and Haizhou Li. 2024b. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*.

Wei Luo, Mengying Xu, and Hanjiang Lai. 2023. Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In *International Conference on Multimedia Modeling*, pages 411–422. Springer.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 164–172.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11205–11214.

Judea Pearl. 2009a. Causal inference in statistics: An overview.

Judea Pearl. 2009b. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press, Cambridge.

Judea Pearl. 2010. An introduction to causal inference. *The international journal of biostatistics*, 6(2).

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

Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 493–502.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. 2022. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55.

Ali Shojaie and Emily B Fox. 2022. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9(1):289–319.

Qiya Song, Bin Sun, and Shutao Li. 2022. Multimodal sparse transformer network for audio-visual speech recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10028–10038.

Xia Song and 1 others. 2024. Quest: Quadruple multimodal contrastive learning with constraints and self-penalization. In *Proceedings of the 2024 Conference on Neural Information Processing Systems (NeurIPS)*.

Mani Kumar Tellamekala, Shahin Amiriparian, Björn W Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. 2023. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414.

Geng Tu, Bin Liang, Ruibin Mao, Min Yang, and Ruifeng Xu. 2023. Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14054–14067.

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L Crowley. 2023. Accommodating missing modalities in time-continuous multimodal emotion recognition. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. 2018. Partial multi-view clustering via consistent gan. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1290–1295. IEEE.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2024. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36.

Zehan Wang, Ziang Zhang, Luping Liu, Yang Zhao, Haifeng Huang, Tao Jin, and Zhou Zhao. 2023. Extending multi-modal contrastive representations. *arXiv preprint arXiv:2310.08884*.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. 2024. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446.

Dingkang Yang, Kun Yang, Haopeng Kuang, Zhaoyu Chen, Yuzheng Wang, and Lihua Zhang. 2024. Towards context-aware emotion recognition debiasing from a causal demystification perspective via deconfounded training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2021. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12996–13010.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. 2024. Unifying causal representation learning with the invariance principle. *arXiv preprint arXiv:2409.02772*.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2023. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539.

Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. 2020. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2402–2415.

Wenzheng Zhang and Karl Stratos. 2021. Understanding hard negatives in noise contrastive estimation. *arXiv preprint arXiv:2104.06245*.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. 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)*, pages 2608–2618.

Q. Zhou and 1 others. 2024. Tcl-map: Token-level contrastive learning with modality-aware prompting. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*.