Enhancing the Quality of Multimodal Sentiment Analysis via High-Quality Affect Enhancement Network

Anonymous submission

Abstract

Multimodal Sentiment Analysis (MSA) is a technique for analysis human sentiment from heterogeneous modality representations. Existing works focus on designing complex deep learning networks such as modal reconstruction networks or fusion networks, aiming to achieve more powerful performance. However, such works are limited to analyzing the relationships between heterogeneous modalities superficially, failing to deeply decouple the co-causal structure of cross-modal non-sentiment confounding noise and its masking effect on MSA, resulting in the tendency to learn spurious correlations when confronted with data bias. In addition, existing modal reconstruction networks are prone to the modal confusion problem, which further harms the accuracy of MSA. To address these challenges, we propose HQAENet, a novel **H**igh-**Q**uality **A**ffect **E**nhancement Network, that deeply explores the expression chain of sentiment cues in multimodal sentiment analysis, and effectively suppresses the propagation of non-sentiment confounding noise in the sentiment cues chain, thus enhancing the model's ability to capture real sentiment expressions. Furthermore, to address the modal confusion problem, we incorporate an Adversarial Reconstruction Network to enhance cross-modal discrimination. As a plug-and-play denoising network, HQAENet can be flexibly integrated into most MSA methods. Comprehensive experiments on multiple benchmarks definitely demonstrate the effectiveness of the proposed network.

Introduction

Multimodal Sentiment Analysis (MSA) is a technique for analyzing speakers' sentiment from three heterogeneous modalities, namely text, visual, and acoustic, and has been widely used in technical domains such as dialog systems (Ghosal et al. 2019), social media analysis (Somandepalli et al. 2021; Stappen et al. 2021), and human-computer interaction (Cambria et al. 2017; Poria et al. 2017). Thanks to the development of multimodal retrieval (Zhang et al. 2021; Wei et al. 2024), researchers have proposed Aggregation-based methods (Hazirbas et al. 2016; Zadeh et al. 2018; Fu et al. 2022), Alignment-based methods (Tsai et al. 2019; Lv et al. 2021), and Reconstruction-based methods (Sun et al. 2023; Li, Wang, and Cui 2023; Li et al. 2024), which have achieved impressive performance. Despite the encouraging performance, such models are limited to analyzing the relationships between heterogeneous modalities superficially, failing to explore the co-causal structure of cross-modal non-sentiment confounding noise and its masking effect on MSA task, resulting in these models being prone to learn spurious correlations when confronted with data bias.

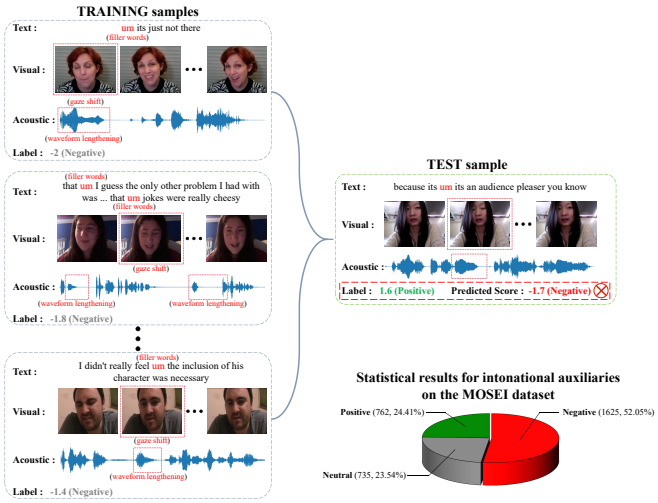

Figure 1: Example of non-sentiment confounding noise harming MSA performance. As they are obvious in all three modalities, they cause a model to learn misleading sentiment associations when confronted with data bias.

Cross-modal non-sentiment confounding noise is mainly caused by non-sentiment factors with distinctive features. Specifically, when speakers are influenced by factors such as environment, expression habits, or prolonged speech, they tend to employ non-sentiment intonational auxiliaries such as "um" or "uh" to aid their expression, to ensure that their speech is back-to-front relevant (Clark and Tree 2002). However, when there is data bias in the training data, most of the current deep learning methods are susceptible to that noise and learn spurious correlations between speakers and their sentiment. As shown in Fig. 1, people subconsciously resort to the intonational auxiliary **"um"** to aid expression when they are thinking or hesitating. Although these intonational auxiliary express distinctive features in all three modalities (**filler words in text**, **gaze shift in visual**, and **waveform lengthening in acoustic**), they do not affect the

speaker's sentiment state (Clark and Tree 2002). However, when there is a training data bias (e.g., the Negative labels occupies 52.05% in the MOSEI dataset), the model tends to capture this notable feature and incorrectly assumes that it expresses negative sentiment, which is refered to the **"masking effect"**. Thanks to Xu et al.(2025), recapitulative causal graph was introduced to decouple the effects of that noise on Multimodal Language Understanding. However, the method fails to explore the co-causal structure of the cross-modal non-sentiment confounding noise, thus failing to eliminate its effect on the model from the causal perspective.

To address these challenges, We first observe the Sentiment Cues Chain that can reflect sentiment relations, to mine the co-causal structure of cross-modal non-sentiment confounding noise and its masking effect on MSA, and accordingly propose a novel High-Quality Affect Enhancement Network (**HQAENet**), which can enhance the model's ability to capture real sentiment representations by suppressing the propagation of non-sentiment confounding noise in the Sentiment Cues Chain without strict identifiability assumptions. As shown in Fig. 2(a), traditionally, the real sentiment representation Z is unobservable despite it is directly associated with sentiment label Y. Accordingly, we require to predict Y by analyzing the semantic representation M of the observation input X. However, both X and M are susceptible to non-sentiment confounding noise N, which produces a masking effect on MSA. Accordingly, inspired by causal paths, we propose HQAENet, which suppresses the propagation of N by learning non-sentiment confusing noise N^{learn}, as shown in Fig. 2(b). Specifically, we analyze the co-causal structure N_{uni} of cross-modal non-sentiment confounding noise and its specific representations in the three modalities. As in Fig. 2(c), for the aforementioned intonational auxiliary noise N_{uni}, its representations in the three modalities are **filler words** N_T, **gaze shifts** N_V, and **waveform lengthening** N_A, respectively. Subsequently, distinct from Visual and Acoustic modalities that are susceptible to inconsistent sentiment expression across time series, Text modality is a token-based sequence with more explicit sentiment tendencies, which is suitable to learn to reconstruct the non-sentiment confounding noise N_T^{learn} from the observations X_T, and then reconstruct its cross-modal noises N_V^{learn} and N_A^{learn} by downscaling in the noise space, achieve the effect of suppressing the propagation of N_{uni}, which enhance model's ability to capture real sentiment representations. Furthermore, following Wen et al.(2025), we apply and improve the Adversarial Reconstruction Network as a plug-and-play module to alleviate the modal confusion problem during cross-modal reconstruction of MSA models.

In summary, the contributions of this work are as follows:

- First observe the negative effect of cross-modal non-sentiment confounding noise on MSA, and reveal its co-causal structure and the masking effect by analyzing the Sentiment Cues Chain.
- Inspired by causal paths, we propose HQAENet, which suppresses the propagation of non-sentiment confounding noise in the Sentiment Cues Chain, and improve the Adversarial Reconstruction Network to mitigate the modal confusion problem.
- Experiments on three datasets show that the proposed HQAENet can significantly improve the performance of existing baselines, and achieve SOTA performance.

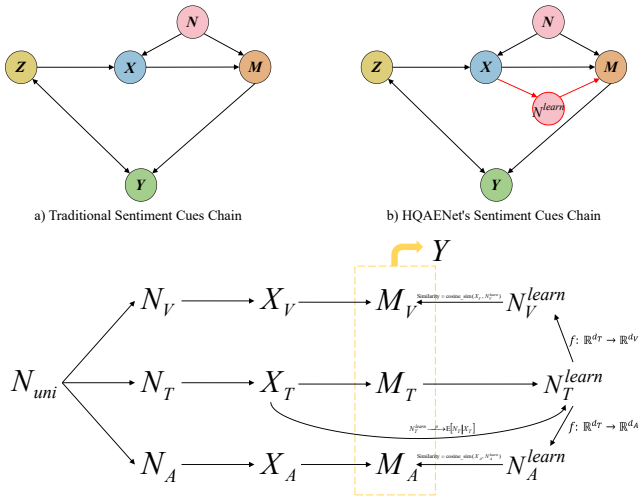

Figure 2: Sentiment Cues Chain with an example, where Z represents the real sentiment representation, X represents the observed input, M represents the semantic representation, Y represents the sentiment label, N represents the non-sentiment confounding noise, and N^{learn} represents the learned non-sentiment confounding noise.

The codes and more experimental details for this paper are available in the Supplementary Materials.

Related Work

Multimodal Sentiment Analysis. Multimodal Sentiment Analysis (MSA) is a technique to analysis and understand varieties of human sentiments by learning the representations from three modalities: text, visual, and acoustic (Morency, Mihalcea, and Doshi 2011). Early on, researchers usually employed Aggregation-based fusion methods (Hazirbas et al. 2016; Ngiam et al. 2011; Zadeh et al. 2018) to integrate multiple modalities into a single modality feature. However, such simple fusion methods will certainly miss the rich cross-modal heterogeneous information. Recent works can be mainly categorized into (a) Alignment-based methods (Tsai et al. 2019; Lv et al. 2021; Yu et al. 2021), which bridge the modality gap and learn effective representations by adaptively aligning potential information across modalities (Ramachandram and Taylor 2017); and (b) Reconstruction-based methods (Li, Wang, and Cui 2023; Sun et al. 2023; Peng et al. 2024), which achieve efficient fusion of unaligned multimodal data by combining modality representations for different modalities reconstruction. Recently, there are HRLF (Li et al. 2024), which reconstructs cross-modal semantic sentiment representations through higher-order representation learning;

SuCI (Xu et al. 2025), which introduced causal interventions into the MSA task and achieved the model's generalization through causal effects; HGAtt-ARN (Wen et al. 2025), which presented the harmful effects of modal confusion on MSA tasks and accordingly proposed an adversarial reconstruction network. Despite the encouraging achievements, these works fail to decouple the co-causal structure of cross-modal non-sentiment confounding noise and its masking effect on MSA, resulting in the tendency to learn spurious correlations when confronted with data bias.

Adversarial Reconstruction Network. Adversarial Reconstruction Network (ARN) mitigates the problem of modal confusion during cross-modal reconstruction by introducing a modal discriminator, and training the model through the adversarial optimization of reconstructor and discriminator. Building on recent advances in multimodal adversarial networks (Peng and Qi 2019; Hu et al. 2019; Yang et al. 2025), HGAtt-ARN (Wen et al. 2025) first presented the Adversarial Reconstruction Network to address the harmful effects of modal confusion on cross-modal reconstruction. In this paper, we improve the ARN network and enable it to be incorporated as a plug-and-play module into most MSA methods.

Methodology

In this section, we will describe the relevant methods of proposed HQAENet. The framework is shown in Fig. 3.

Sentiment Cues Chain in MSA Tasks

To systematically analyze the non-sentiment confounding noise present in the MSA task, we first observed the Sentiment Cues Chain that can visualize the sentiment relations. As shown in Fig. 2, specific sentiment cues are as follows.

► **Link** $Z \rightarrow X \leftarrow N$. Although the real sentiment representation Z is unobservable, in Sentiment Cues Chain, Z can be characterized by observing the input X to express its sentiment, i.e., $Z \rightarrow X$. Similarly, the non-sentiment confounding noise N can affect the real human sentiment by influencing X, i.e., $N \rightarrow X$.

► **Link** $X \rightarrow M \leftarrow N$. M denotes the multimodal semantic representation extracted by the MSA model, which serves as a mediator before the final classifier. The link $X \rightarrow M$ denotes the generic multimodal feature from X. The link $N \rightarrow M$ denotes that non-sentiment confounding noise affects the semantic representation of sentiment, thus generating spurious semantic correlations, an intuitive example of which is the cross-modal noise generated by sentiment-independent intonational auxiliaries in Fig. 1.

► **Link** $Z \rightarrow Y \leftarrow M$. Y denotes sentiment labels. The link $Z \rightarrow Y$ represents that Y has a direct relation to the real sentiment Z. However, Z is unobservable. Consequently, we should estimate the real sentiment by linking $M \rightarrow Y$.

► **Link** $X \rightarrow N^{learn} \rightarrow M$. Figure 2(b) represents the Sentiment Cues Chain of the proposed HQAENet, differs from the traditional ones, inspired by causal paths, HQAENet isolates the propagation path of the non-sentiment confounding noise by linking $X \rightarrow N^{learn} \rightarrow M$, thus enhancing the model's ability to capture the real sentiment representation Z.

Pre-Encoder

The target of MSA task is to predict the speaker's sentiment intensity label Y based on the cross-modal representations X_V, X_T and X_A.Following previous works (Yuan et al. 2021; Sun et al. 2023), we use the powerful BERT Encoder (Kenton and Toutanova 2019) to pre-encode the observation inputs X_T, the LSTM Encoder (Hochreiter 1997) to pre-encode X_V and X_A , which are computed as follows:

$$C_{\{V,A\}} = \mathrm{LSTM}(X_{\{V,A\}}) \tag{1}$$

$$C_T = \mathrm{BertTextEncoder}(X_T) \tag{2}$$

where $C_{\{V,T,A\}}$ represents the Pre-encoding inputs.

High-Quality Affect Enhancement Network

To suppress the non-sentiment confounding noise in the pipeline model's input, we propose HQAENet based on Sentiment Cues Chain, which enhances the input features of each modality during upstream task. HQAENet consists of four steps: generating noise, learning noise, enhancing features, and comparing features.

Generating Noise. From Fig.2, the non-sentiment confounding noise N is unobservable. Thus, we introduce a proxy Noise Generator to produce learnable perturbations that approximate non-sentiment confounding noise in each modality, which is calculated as follows:

$$N'_{\{V,T,A\}} = C_{\{V,T,A\}} + w_{\{V,T,A\}} \cdot \varepsilon \tag{3}$$

where $\varepsilon \sim (0, \sigma^2)$, $N'_{\{V,T,A\}}$ represents the feature with simulated noise for three modalities respectively, and $w_{\{V,T,A\}}$ represents the weight for constructing the corresponding modality noise $N_{\{V,T,A\}}$.

Learning Noise. Given that Text modality is a token-based sequence, which has a more explicit sentiment tendency compared to Visual and Acoustic that are susceptible to time series. Consequently, thanks to the powerful denoising capability of the Bert encoder (Kenton and Toutanova 2019) and the U-Net Network (Ronneberger, Fischer, and Brox 2015), we learn to reconstruct the non-sentiment confounding noise N_T^{learn} by the Asymptotic Consistency approach based on the Text Pre-encoding input C_T, and simultaneously reconstruct its learned cross-modal noise N_V^{learn} and N_A^{learn} by downscaling in the noise space, which are calculated as follows:

$$M_T = \text{U-Net}(N'_T) \tag{4}$$

$$N_T^{learn} = M_T - N'_T \tag{5}$$

where for Visual and Acoustic modalities, there have $f : N_T^{learn} \in \mathbb{R}^{d_T} \rightarrow W_{\{V,A\}} \cdot N_{\{V,A\}}^{learn} \in \mathbb{R}^{\{d_V, d_A\}}$, $W_{\{V,A\}}$ is the learnable parameter and $d_T > \{d_V, d_A\}$.

To maximally simulate the non-sentiment confounding noise, we learn it using a minimized MSE loss:

$$\mathcal{L}^{\mathrm{HQAENet}} = \mathbb{E}[\|\text{U-Net}_\theta(N'_T) - C_T\|_2^2] \tag{6}$$

where θ represents the parameters of the U-Net network. The Asymptotic Consistency theory guarantees unobservable noise N_T when the training data is large enough: $N_T^{learn} \xrightarrow{p} \mathbb{E}[N_T|X_T]$.

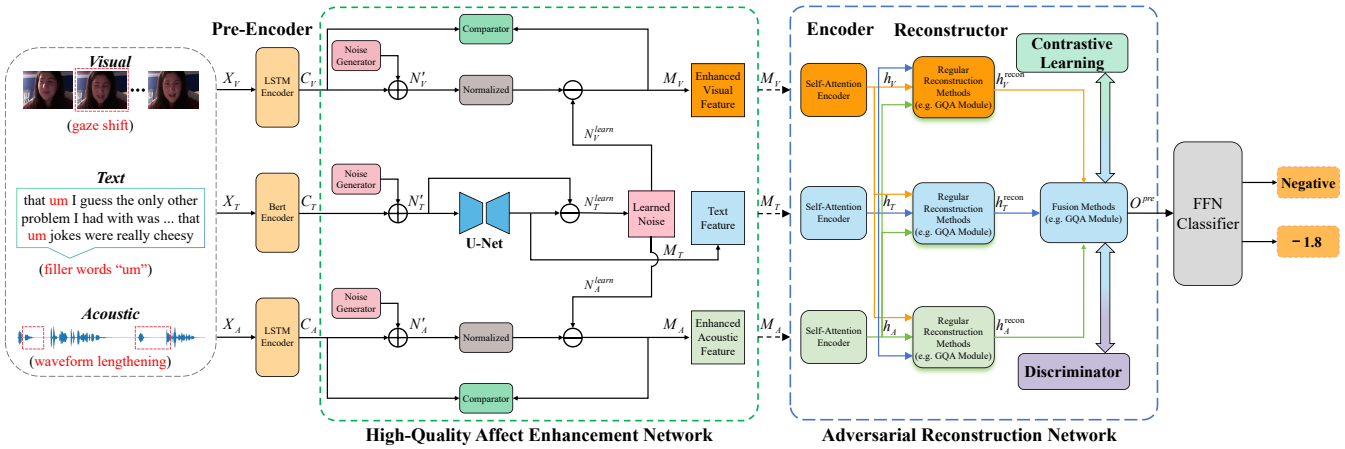

Figure 3: Our proposed High-Quality Affect Enhancement Network (**HQAENet**) integrated in the universal MSA models, which can be easily integrated into extensive ARN models for suppressing non-sentiment confounding noise in upstream task, and achieving the enhancement of real sentiment features.

Enhancing Features. Given that the LSTM Encoder is limited to extracting the contextual information of Visual and Acoustic modalities in the time series, it cannot deal with the non-sentiment confounding noise N_V and N_A that intermixes with the modality features. Accordingly, based on the above theories, we suppress the propagation of N_V and N_A in the Sentiment Cues Chain by HQAENet to obtain high-quality sentiment features, which are calculated as:

$$M_{\{V,A\}} = W_M \cdot \text{Normalized}(N'_{\{V,A\}}) - W_N \cdot N^{learn}_{\{V,A\}} \quad (7)$$

where W_M and W_N are the learnable parameters. $\text{Normalized}(\cdot)$ is the Layer Normalization operation (Ba, Kiros, and Hinton 2016).

Comparing Features. To ensure that the learned noises N^{learn}_V and N^{learn}_A are strongly correlated with their modality, we present the Comparator by comparing the similarity of the enhancement feature $M_{\{V,A\}}$ with the input $C_{\{V,A\}}$, ensuring that $M_{\{V,A\}}$ is reliable with respect to the original modality. Here, we adopt the cosine similarity measure:

$$\text{Sim}(M_{\{V,A\}}, C_{\{V,A\}}) = \frac{M_{\{V,A\}} \cdot C_{\{V,A\}}}{|M_{\{V,A\}}|\,|C_{\{V,A\}}|} \quad (8)$$

Then, we introduce the comparison parameter α, when the similarity of both Visual and Acoustic modalities is greater than α, $M_{\{V,T,A\}}$ is served as input features to the Adversarial Reconstruction Network.

Adversarial Reconstruction Network

To alleviate the modal confusion problem of current reconstruction-based models, based on Wen et al.(2025), we reconstructed the ARN network to adapt arbitrary reconstruction-based models as a plug-and-play method.

Encoder. In this work, we introduce a Self-Attention Encoder (Vaswani et al. 2017) to encode the enhanced features $M_{\{V,T,A\}}$ from HQAENet, which is computed as follows:

$$h_{\{V,T,A\}} = \text{Self-Attention}(M_{\{V,T,A\}}) \quad (9)$$

Reconstructor. The role of Reconstructor is to reconstruct the aggregated features of the three modalities to enhance their modal representations through supervised learning. In this work, based on the latest Grouped Query Attention (GQA) mechanism (Ainslie et al. 2023), we design the GQA Reconstructor to achieve better quality while maintaining high efficiency, which is calculated as follows:

$$\text{Attention}(Q_i, K_g, V_g) = \text{softmax}((\frac{Q_i K_g^T}{\sqrt{d^k}})V_g) \quad (10)$$

$$\begin{aligned}\text{GQA}(Q,K,V) = \text{Concat}(\,&\text{Attention}(Q_1, K_g, V_g), \ldots, \\ &\text{Attention}(Q_h, K_g, V_g))\end{aligned} \quad (11)$$

where $K_g = W^g_K K$, $V_g = W^g_V V$, h is the number of attention heads and g is the number of groups.

For GQA Reconstructor, the formula is as follows:

$$h^{recon}_{\{V,T,A\}} = \text{GQA}(h_{\{V,T,A\}}, h_{\{T,A,V\}}, h_{\{A,V,T\}}) \quad (12)$$

Discriminator. The Discriminator is a core component of the Adversarial Reconstruction Network that acts as a modal discriminator to ensure that the modality reconstructed through Reconstructor does not create the modal confusion problem. The Discriminator consists of multiple Linear Layers and an Activation Layer, and predicts a scalar between 0 and 1 for each time series by Softmax normalization, indicating which modality it belongs to (i.e., 0 for Visual, 0.5 for Text, and 1 for Acoustic)(Hu et al. 2023; Wen et al. 2025):

$$\mathbf{D}_i = \mathbf{D}(h^{recon}_i) \in \mathbb{R}^{T\times 1} \quad (13)$$

where T represents the time series, and $i = \{V, T, A\}$ represents the Visual, text and acoustic modalities, respectively.

Accordingly, the ARN Network's Loss is calculated as:

$$\mathcal{L}^{\text{ARN}} = \mathbb{E}_h[\log(\mathbf{D}_A(1-\mathbf{D}_V))] + \mathbb{E}_h[-\log(\mathbf{D}_T(1-\mathbf{D}_T))] \quad (14)$$

where $\mathbb{E}$ denotes the expectation in the current Batch Size for all time series in sample group $h = (h^{recon}_V, h^{recon}_T, h^{recon}_A)$.

Contrastive Learning. For a set of samples $h^{recon}_{\{V,T,A\}} = (h^j_V, h^j_T, h^j_A)$ from reconstructor, Multimodal Contrastive Learning can map samples that are similar to the Text feature to neighboring locations to learn stronger sentiment features, and map different samples to relatively distant locations (Wei et al. 2024), with the loss function computed as:

$$\mathcal{L}^{\mathrm{CL}} = -\sum_{j=1}^{n} \log \frac{e^{[\mathrm{sim}(h^j_V, h^j_T) + \mathrm{sim}(h^j_A, h^j_T)]/\tau}}{\prod_{i\in\{V,A\}} [\sum_{k=1}^{N} e^{\mathrm{sim}(h^k_i, h^j_T)/\tau} + e^{\mathrm{sim}(h^j_i, h^j_T)/\tau}]} \tag{15}$$

where $\mathrm{sim}(\cdot)$ is the similarity calculation function in Eq. 8, τ is the scalar temperature parameter, N is the sample size.

FFN Classifier

At the classification layer, we first aggregate the three modality reconstruction features, and then predict the speaker's sentiment label $\hat{Y}$ through a Feed-Forward Neural network, which is calculated as follows:

$$O^{pre} = \mathrm{GQA}(h^{recon}_V, h^{recon}_T, h^{recon}_A) \tag{16}$$

$$\hat{Y} = \mathrm{FFN}(O^{pre}) \tag{17}$$

Joint Loss Function

We adopt a joint training scheme, constructing Reconstruct_Loss (Eqs.18), HQAENet_Loss (Eqs.6), ARN_Loss (Eqs.14), Contrastive_Loss (Eqs.15) and Predict_Loss (Eqs.19), and a joint loss function is used to learn the parameters (Eqs.20) with the learning objective of minimizing the loss between the model's predicted sentiment label $\hat{Y}$ and the real sentiment label Y.

Reconstruct Loss. Following Sun et al.(2023), we utilize low-resource feature reconstruction to encourage the model to learn modal features that reflect the real sentiment label Y, and use $\mathrm{Smooth}_{\mathrm{L1}}$ loss (Yuan et al. 2021) to evaluate the reconstruction quality, which is calculated as:

$$\mathcal{L}^{recon} = \sum_{i=1}^{\{V,T,A\}} \mathrm{Smooth}_{\mathrm{L1}}(h^{recon}_i, Y) \tag{18}$$

Predict Loss. Following previous work (Yuan et al. 2021; Yu et al. 2021), we employ the $L1$ loss as the predicted loss, which is calculated as follows:

$$\mathcal{L}^{pre} = \frac{1}{N}\sum_{j=1}^{N}(\left|\hat{Y}_j - Y_j\right|) \tag{19}$$

Joint Loss. Finally, the Joint Loss is calculated as:

$$\mathcal{L} = \mathcal{L}^{pre} + \mu_1\mathcal{L}^{\mathrm{HQAENet}} + \mu_2\mathcal{L}^{recon} + \mu_3\mathcal{L}^{\mathrm{ARN}} + \mu_4\mathcal{L}^{\mathrm{CL}} \tag{20}$$

where μ_1,μ_2, μ_3 and μ_4 denote the hyperparameters that balance the contribution of each Loss to the Joint Loss.

Experiment

Datasets and Implementation Details

We conduct experiments on three datasets, **CMU-MOSI**, **CMU-MOSEI** and **CH-SIMS**, to evaluate the effectiveness of HQAENet. The **CMU-MOSI** includes speakers' views on a wide range of movie topics collected from YouTube, which consists of 93 videos from 89 speakers with 2,199 discourse-level video clips, and each video clip is labeled with a sentiment intensity from -3 (Highly Negative) to 3 (Highly Positive). The **CMU-MOSEI** is a modified version of MOSI with a larger training sample and richer topics, which contains 22,856 discourse-level video clips from 1,000 speakers for 250 different topics. And the **CH-SIMS** collected 2,281 voice-level video clips from 60 videos including movies, TV shows, and variety shows, and labeled with sentiment intensities from -1 (Negative) to 1 (Positive).

Our experimental environment includes PyTorch 1.8.0 and CUDA 11.1, setting $\mathrm{Batch\_Size}$ to 16, Bert's learning rate to $2e$-5, and other networks' learning rates to $1e$-4. We train HQAENet on NVIDIA RTX 3090 GPUs. And multiple experiments are conducted with extensive random seeds to select the model parameter that performs best on the validation set and evaluate it on the test set.

Baseline Models

To validate the effectiveness of HQAENet, we compared the performance improvement of the HQAENet plugin on five representative models. These models include Aggregation-based MISA (Hazarika, Zimmermann, and Poria 2020), Alignment-based Self-MM (Yu et al. 2021), recent Reconstruction-based DMD (Li, Wang, and Cui 2023), HGAtt-ARN (Wen et al. 2025), and the GQA-ARN proposed in this paper. Additionally, we compared the improvements of SuCI (Xu et al. 2025) and HQAENet. We tuned all models to their optimal performance for a fair comparison.

Comparison with State-of-the-art Models

The target of the MSA task is to predict speaker's sentiment label $\hat{Y}$ based on their cross-modal representations. Following previous studies (Li, Wang, and Cui 2023; Sun et al. 2023), we use 7-class accuracy (*Acc7*), 2-class accuracy (*Acc2*), and *F1* score as the main evaluation metrics for MOSI and MOSEI, and perform separate Aligned and Unaligned setting comparison experiments (Li, Wang, and Cui 2023). While on CH-SIMS, following Yu et al.(2020), we report *Acc5*, *Acc2* and *F1* score. We compare the HQAENet-based and SuCI-based methods with a wide range of SOTA methods proposed in recent years, including NHFNet (Fu et al. 2022), EMT-DLFR (Sun et al. 2023), MPLMM (Guo, Jin, and Zhao 2024) and HRLF (Li et al. 2024).

Table 1 presents the baseline comparison on three datasets under the Aligned setting (Unaligned setting can be found in Appendix). From Tab.1, we can observe that: (1) The proposed HQAENet can significantly enhance the performance of the baseline models. In particular, for MISA (Hazarika, Zimmermann, and Poria 2020), HQAENet can achieve an average performance enhancement of **1.39%** . This indicates that the proposed HQAENet can significantly enhance the performance of most MSA models in a model-independent manner. (2) Compared to the SuCI method (Xu et al. 2025), HQAENet provides a more significant improvement to baselines' performance, suggesting that compared to SuCI that

Methods	MOSI			MOSEI			CH-SIMS			Average Improve
	Acc7(%)	Acc2(%)	F1(%)	Acc7(%)	Acc2(%)	F1(%)	Acc5(%)	Acc2(%)	F1(%)	
NHFNet$_{2022}$	42.9	83.4	83.5	53.1	85.0	84.9	42.5	78.1	78.3	-
EMT-DLFR$_{2023}$	46.0	85.0	85.0	54.5	86.0	86.0	43.5	79.9	80.1	-
MPLMM$_{2024}$	46.1	85.7	85.6	54.2	85.8	85.8	43.9	80.1	80.3	-
HRLF$_{2024}$	45.9	85.3	85.2	54.1	85.5	85.5	43.1	79.6	79.7	-
MISA$^{\dagger}_{2020}$	40.2	81.8	81.9	51.3	82.3	82.3	-	76.4	76.6	-
MISA+SuCI$^{\dagger}_{2025}$	$41.6^{+1.4}$	$83.3^{+1.5}$	$83.1^{+1.2}$	$52.6^{+1.3}$	$83.5^{+1.2}$	$83.2^{+0.9}$	-	$77.0^{+0.6}$	$77.2^{+0.6}$	↑1.08
MISA+**HQAENet**	$41.8^{+1.6}$	$83.6^{+1.8}$	$83.5^{+1.6}$	$52.5^{+1.2}$	$83.8^{+1.5}$	$83.7^{+1.4}$	-	$77.3^{+0.9}$	$77.8^{+1.2}$	↑**1.39**
Self-MM$^{\dagger}_{2021}$	41.6	83.9	84.1	52.9	83.9	83.8	43.1	78.6	78.6	-
Self-MM+SuCI$^{\dagger}_{2025}$	$42.0^{+0.4}$	$84.3^{+0.4}$	$84.6^{+0.5}$	$53.6^{+0.7}$	$84.2^{+0.3}$	$84.2^{+0.4}$	$43.4^{+0.3}$	$79.2^{+0.6}$	$79.4^{+0.8}$	↑0.49
Self-MM+**HQAENet**	$42.2^{+0.6}$	$84.6^{+0.7}$	$84.7^{+0.6}$	$53.5^{+0.6}$	$84.9^{+1.0}$	$84.7^{+0.9}$	$43.6^{+0.5}$	$79.7^{+1.1}$	$79.5^{+0.9}$	↑1.15
DMD$^{\dagger}_{2023}$	41.0	83.3	83.2	53.5	84.1	84.0	43.2	80.0	79.9	-
DMD+SuCI$^{\dagger}_{2025}$	$42.2^{+1.2}$	$84.6^{+1.3}$	$84.5^{+1.3}$	$\mathbf{54.6}^{+1.1}$	$85.8^{+1.7}$	$85.7^{+1.7}$	$43.4^{+0.2}$	79.8	79.9	↑0.92
DMD+**HQAENet**	$42.0^{+1.0}$	$84.6^{+1.3}$	$84.7^{+1.5}$	$54.5^{+1.0}$	$85.5^{+1.4}$	$85.5^{+1.5}$	$43.7^{+0.5}$	$80.1^{+0.1}$	$80.0^{+0.1}$	↑0.88
HGAtt-ARN$_{2025}$	45.4	85.8	85.7	53.4	85.6	85.4	44.0	80.7	80.5	-
HGAtt-ARN+SuCI	$45.7^{+0.3}$	$86.1^{+0.3}$	$85.9^{+0.2}$	$53.5^{+0.1}$	85.6	$85.5^{+0.1}$	43.9	80.5	80.4	↑0.06
HGAtt-ARN+**HQAENet**	$46.2^{+0.8}$	$\mathbf{86.6}^{+0.8}$	$\mathbf{86.6}^{+1.1}$	$53.9^{+0.5}$	$86.2^{+0.6}$	$85.9^{+0.5}$	$\mathbf{44.2}^{+0.2}$	$\mathbf{80.9}^{+0.2}$	$\mathbf{80.6}^{+0.1}$	↑0.53
GQA-ARN	45.9	85.5	85.4	53.2	85.3	85.5	43.1	80.1	80.0	-
GQA-ARN+SuCI	$46.1^{+0.2}$	$85.7^{+0.2}$	$85.5^{+0.1}$	$53.5^{+0.3}$	$85.9^{+0.6}$	$86.0^{+0.5}$	$43.2^{+0.1}$	$80.4^{+0.3}$	$80.2^{+0.2}$	↑0.27
GQA-ARN+HQAENet	$\mathbf{47.4}^{+1.5}$	$86.4^{+0.9}$	$86.3^{+0.9}$	$53.6^{+0.4}$	$\mathbf{86.5}^{+1.2}$	$\mathbf{86.4}^{+0.9}$	$43.7^{+0.6}$	$80.5^{+0.4}$	$80.7^{+0.7}$	↑0.83

Table 1: Main results of the baseline models on MOSI and MOSEI, where the marker "†" represents the results from Xu et al.(2025), the best results are in **bolded** and the sub-optimal are underlined.

relies on strong preconditions, HQAENet is able to suppress most non-sentiment confounding noise flexibly, thus significantly enhancing the baselines' performance. (3) For both the latest HGAtt-ARN method (Wen et al. 2025) and the GQA-ARN method adopted in this work, HQAENet improves them to the newest SOTA or sub-optimal performance. In particular, the Acc2 metric achieves a performance of 86.6% and 80.9% on the MOSI and CH-SIMS datasets, respectively. This indicates that HQAENet can effectively suppress non-sentiment confounding noise via Sentiment Cues Chain and provide high quality affective representations for models, thus effectively improving the performance of existing SOTA models.

Ablation Study

To observe the efficiency of the different modal enhancements in HQAENet in terms of performance improvement, and the effect of different components in the ARN network, we conduct comprehensive ablation studies on three datasets utilizing *F1* score, as shown in Tab. 2.

Effect on High-Quality Affect Enhancement Network. In this experiment, we ablated the Visual Enhanced Feature and Acoustic Enhanced Feature in HQAENet to observe the efficiency of the different modal enhancements. From the results of Tab.2, it can be observed that methods w/o Visual Enhanced and w/o Acoustic Enhanced significantly decrease in all metrics compared to the original method. In particular, w/o Visual Enhanced shows an average decrease of **0.7%**, which suggest that HQAENet can effectively suppress the propagation of non-sentiment confounding noise in the Sentiment Cues Chain, especially for Visual modality, which significantly enhances its real sentiment representation.

Methods	MOSI	MOSEI	CH-SIMS	Average Drop
GQA-ARN+HQAENet	**86.3**	**86.4**	**80.7**	-
w/o Visual Enhanced	$85.0^{\downarrow 1.3}$	$85.8^{\downarrow 0.6}$	$80.4^{\downarrow 0.3}$	0.7
w/o Acoustic Enhanced	$85.7^{\downarrow 0.6}$	$85.8^{\downarrow 0.6}$	$80.5^{\downarrow 0.2}$	0.4
w/o Discriminator	$85.8^{\downarrow 0.5}$	$85.6^{\downarrow 0.8}$	$80.3^{\downarrow 0.4}$	0.5
w/o Contrastive Learning	$86.1^{\downarrow 0.2}$	$86.0^{\downarrow 0.4}$	$80.5^{\downarrow 0.2}$	0.2
w/o h_V^{recon}	$85.7^{\downarrow 0.6}$	$85.9^{\downarrow 0.5}$	$80.3^{\downarrow 0.4}$	0.5
w/o h_T^{recon}	$84.5^{\downarrow 1.8}$	$84.1^{\downarrow 2.3}$	$79.6^{\downarrow 1.1}$	**1.7**
w/o h_A^{recon}	$85.5^{\downarrow 0.8}$	$85.7^{\downarrow 0.7}$	$80.4^{\downarrow 0.3}$	0.6

Table 2: Ablation study on three datasets, where "w/o" means without that component.

Effect on Adversarial Reconstruction Network. To verify the effect of ARN Network on performance, we ablated two of the important components: Discriminator and Contrastive Learning module, as in Tab.2. Experimental results indicate that Discriminator is able to solve the modal confusion problem of the Reconstruction-based methods well in ARN networks, thus improving the model's performance, whereas the Contrastive Learning module seems to have little effect on the performance enhancement of the model.

Effect on Different Modalities. To observe the effect of different modalities on the model performance, we ablate the reconstructed modalities h_V^{recon}, h_T^{recon}, and h_A^{recon}, respectively, as shown in Tab.2. And we can observe that all three modal information contributes significantly to the model performance. In particular, w/o h_T^{recon} significantly shows an average reduction of **1.7%**. From the above results, we can speculate that the heterogeneous modal information has distinctive contribution to MSA task. Moreover, the Text modality based on token sequences has a more explicit sentiment tendency compared to Visual and Acoustic, and has a

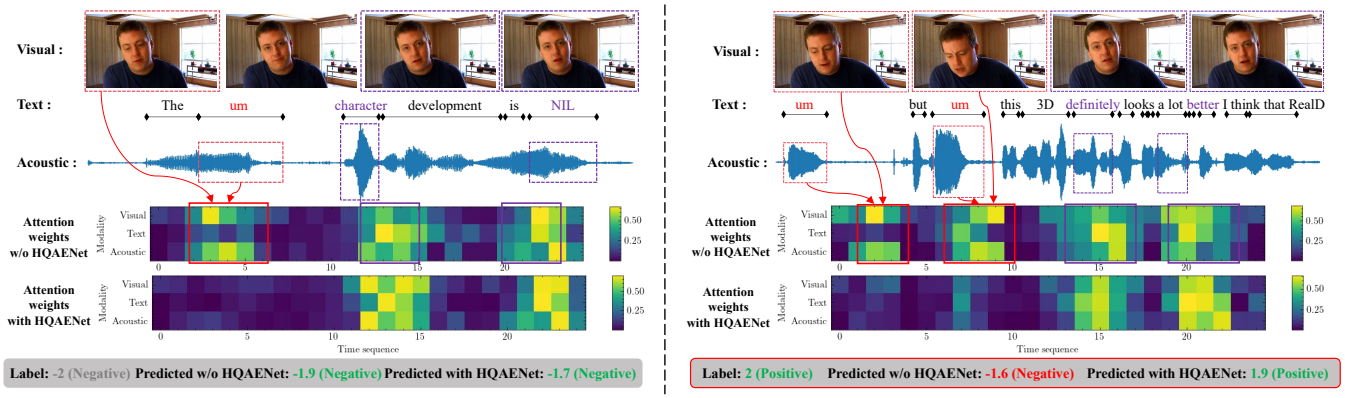

Figure 4: Visualization of self-attention weights on sample "LSi-o-IrDMs_10"(left) and "LSi-o-IrDMs_14"(right). Meaningful attention areas are highlighted by colored rectangles. Red boxes represent signals with non-sentiment confounding noise and purple boxes represent meaningful signals.

more significant impact on MSA task.

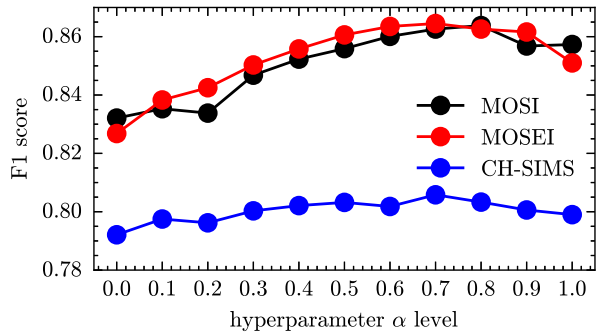

Figure 5: The effect of different level of hyperparameter α.

Sensitivity of the Hyperparameter Level.

In this study, we conducted a sensitivity analysis of the key comparison parameter α to explore its impact on model performance. α plays a crucial role in HQAENet by evaluating the similarity between the enhancement feature M and the original modality C, which in turn ensures that M is reliable with respect to the original modality. Specifically, when the similarity of both visual and acoustic modalities is higher than the threshold α, the feature M will be inputted into the ARN network, and the experimental results are shown in Fig. 5. The results show that the *F1* score shows a tendency of rising and then stabilizing with the increasing of α, and reaches its optimum between α of about 0.6 and 0.8, after which the *F1* score slightly decreases upon further increase of α. This suggests that within a certain range, increasing the α can suppress non-sentiment confounding noise while ensuring that M is reliable on the original modality, thus improving the quality of sentiment features. However, when α is too high, it may cause the model to be too strict, and introducing noise corruption into M, causing the HQAENet denoising to fail, thus harming the model performance.

Case Study

To observe HQAENet's ability to suppress non-sentiment confounding noise such as intonational auxiliaries, we visualize the self-attention weights to study the signals captured by HQAENet, as shown in Fig. 4. Following Sun et al.(2023), we randomly selected samples "LSi-o-IrDMs" from MOSI dataset, and halve the acoustic and visual sequences using average pooling with a stride of 2.

From Fig. 4, we can observe that: (1) Compared to the model w/o HQAENet, model with HQAENet can more effectively focus on sentiment-related signals while suppressing non-sentiment confounding noise. (2) Compared to Visual and Acoustic signals encoded by LSTM, the Text signals encoded by Bert can more accurately focus on sentiment-related signals, which provides strong experimental support for the denoising direction of HQAENet. (3) When dealing with the challenging sample "LSi-o-IrDMs_14", the model w/o HQAENet is susceptible to non-sentiment confounding noise and incorrectly predicts sentiment labels as -1.6 (**Negative**), whereas model with HQAENet suppresses that noise well and predicts a **Positive** sentiment score of 1.9, which comes pretty close to the ground-truth label 2.0. These interesting observations qualitatively reveal the negative effect of non-sentiment confounding noise on MSA task, and further demonstrates the proposed HQAENet's interpretability and effectiveness.

Conclusion

This work first observes the masking effect of cross-modal non-sentiment confounding noise on MSA task, and accordingly proposes HQAENet to suppress its propagation in Sentiment Cues Chain, enhances the model's ability to capture the real sentiment signals when confronted with data bias. Furthermore, we incorporate an adversarial reconstruction network to address the modal confusion problem caused by modal reconstruction. As a plug-and-play denoising network, HQAENet can be widely applied to most MSA models, and experiments on several benchmarks demonstrate the effectiveness of the proposed HQAENet.

References

Ainslie, J.; Lee-Thorp, J.; de Jong, M.; Zemlyanskiy, Y.; Lebron, F.; and Sanghai, S. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4895–4901.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Cambria, E.; Das, D.; Bandyopadhyay, S.; and Feraco, A. 2017. Affective computing and sentiment analysis. *A practical guide to sentiment analysis*, 1–10.

Clark, H. H.; and Tree, J. E. F. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1): 73–111.

Fu, Z.; Liu, F.; Xu, Q.; Qi, J.; Fu, X.; Zhou, A.; and Li, Z. 2022. NHFNET: A non-homogeneous fusion network for multimodal sentiment analysis. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 154–164.

Guo, Z.; Jin, T.; and Zhao, Z. 2024. Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1726–1736.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

Hazirbas, C.; Ma, L.; Domokos, C.; and Cremers, D. 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, 213–228. Springer.

Hochreiter, S. 1997. Long Short-term Memory. *Neural Computation MIT-Press*.

Hu, P.; Peng, D.; Wang, X.; and Xiang, Y. 2019. Multimodal adversarial network for cross-modal retrieval. *Knowledge-Based Systems*, 180: 38–50.

Hu, Y.; Chen, C.; Li, R.; Zou, H.; and Chng, E. S. 2023. MIR-GAN: Refining Frame-Level Modality-Invariant Representations with Adversarial Network for Audio-Visual Speech Recognition. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Li, M.; Yang, D.; Liu, Y.; Wang, S.; Chen, J.; Wang, S.; Wei, J.; Jiang, Y.; Xu, Q.; Hou, X.; et al. 2024. Toward Robust Incomplete Multimodal Sentiment Analysis via Hierarchical Representation Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6631–6640.

Lv, F.; Chen, X.; Huang, Y.; Duan, L.; and Lin, G. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2554–2562.

Morency, L.-P.; Mihalcea, R.; and Doshi, P. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, 169–176.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Y.; et al. 2011. Multimodal deep learning. In *ICML*, volume 11, 689–696.

Peng, C.; Chen, K.; Shou, L.; and Chen, G. 2024. CARAT: Contrastive Feature Reconstruction and Aggregation for Multi-Modal Multi-Label Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14581–14589.

Peng, Y.; and Qi, J. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1): 1–24.

Poria, S.; Cambria, E.; Bajpai, R.; and Hussain, A. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37: 98–125.

Ramachandram, D.; and Taylor, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6): 96–108.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Somandepalli, K.; Guha, T.; Martinez, V. R.; Kumar, N.; Adam, H.; and Narayanan, S. 2021. Computational media intelligence: Human-centered machine analysis of media. *Proceedings of the IEEE*, 109(5): 891–910.

Stappen, L.; Baird, A.; Schumann, L.; and Schuller, B. 2021. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing*, 14(2): 1334–1350.

Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1): 309–325.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Wei, P.; Ouyang, H.; Hu, Q.; Zeng, B.; Feng, G.; and Wen, Q. 2024. VEC-MNER: Hybrid Transformer with Visual-Enhanced Cross-Modal Multi-level Interaction for Multimodal NER. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 469–477.

Wen, Q.; Wei, P.; Li, F.; Hu, Q.; Zeng, B.; and Feng, G. 2025. HGAtt-ARN: A Novel Adversarial Reconstruction Network Based on Higher-order Gate Attention for Incomplete Multimodal Sentiment Analysis. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 1479–1487.

Xu, Z.; Yang, D.; Li, M.; Wang, Y.; Chen, Z.; Chen, J.; Wei, J.; and Zhang, L. 2025. Debiased multimodal understanding for human language sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14450–14458.

Yang, B.; Xiang, X.; Kong, W.; Zhang, J.; and Yao, J. 2025. SF-GAN: Semantic fusion generative adversarial networks for text-to-image synthesis. *Expert Systems with Applications*, 262: 125583.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 3718–3727.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10790–10797.

Yuan, Z.; Li, W.; Xu, H.; and Yu, W. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4400–4407.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L.-P. 2018. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhang, W.; Lin, H.; Han, X.; and Sun, L. 2021. De-biasing Distantly Supervised Named Entity Recognition via Causal Intervention. 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)*, 4803–4813.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) yes

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) yes

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) yes

2.4. Proofs of all novel claims are included (yes/partial/no) partial

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) yes

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) yes

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) yes

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) NA

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) partial

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) partial

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/-partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) partial

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) partial

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) partial

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) partial