Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 16, No. 2, 2026

Noise-Aware Causal Orthogonal Gating Network for Robust and Interpretable Multimodal Sentiment Analysis

Junrui Li [1], Jing Chen [1], Chang Li [2], Wei Zhang [1]

[1] School of Physics and Optoelectronic Engineering, Guangdong University of Technology, Guangzhou, Guangdong, China
[2] School of Instrumental Science and Optoelectronic Engineering, Hefei University of Technology, Hefei, Anhui, China

Abstract: Across online social platforms, the sheer expansion of multimodal content has made Multimodal Sentiment Analysis (MSA) both indispensable and technically difficult within computational social systems. Cross-modal heterogeneity is the core of the whole problem: textual, visual, and acoustic data streams almost never align completely in a clean, simple way. On top of this basic mismatch, things like noise, missing types of modality, and wrong inter-modal connections end up hurting both how reliable existing models are, and also how easy they are to understand. Given all these issues, this study puts forward the Noise-Aware Causal Orthogonal Gating Network (NCOG), which is a new framework made to add causal inference theory into adaptive multimodal fusion work. It does not use the traditional unimodal nonlinear gating that most methods rely on; instead, NCOG splits the gating work into two separate submodules that are orthogonal but can work together to complement each other. One of these submodules is called the Reliability Gate, and it filters signals using clear noise proxies — things like inter-frame variation, pitch energy fluctuation, and ASR confidence. The other one, named the Causal Contribution Gate, does not work the same way as this front-end screening step. It uses timestep-level counterfactual masking to work out the real causal effects between different modalities, and at the same time it gets rid of spurious correlations that do not mean anything. Along with this two-gate design, there is also a cross-modal attention module that dynamically adjusts how much weight each modality gets. When the input data has noise or some parts are missing, this module also makes the model more stable and harder to break. When we tested it on real data, NCOG-MSA got 88.12 percent accuracy, 87.11 percent F1 score, 0.681 for MAE, and 0.809 for PCC on the CMU-MOSI dataset. For the CMU-MOSEI dataset, the corresponding numbers are 87.44 percent accuracy, 89.04 percent F1, 0.483 MAE, and 0.816 PCC. If we compare it with the best existing fusion baselines, which include ACMG, TMFN, and HCAN, our framework shows better results on every single one of these evaluation metrics. Putting all these test results together, we can see that combining explicit noise modeling and causal reasoning makes multimodal sentiment analysis more robust, easier to understand, and works better on new data. This also gives a solid basic framework for building affective intelligence that can work in complex real-world social systems.

Keywords: Multimodal Sentiment Analysis; Signal-to-Noise Causal Orthogonality; Gating Mechanism; Deep Learning.

1. Introduction

Emotion is right at the center of how humans communicate and make choices. It shapes how we think, how we act, and how we interact with each other, and these effects are not random accidents, nor can they be turned into simple formulas easily. As social media, video sharing platforms and connected communication systems have grown a lot in recent years, the ways we show our feelings — and what's just as important, the different shapes this expression takes — have changed in big basic ways. Emotional sharing is no longer stuck to just writing words down, these days it happens through layered, detailed emotional hints that spread over many different types of content, including written words, spoken voice, and visual signals. As this multimodal configuration has become increasingly commonplace, traditional single-modality sentiment analysis methods have run into difficulties that were far less visible at earlier scales; under such conditions, Multimodal Sentiment Analysis (MSA) has accordingly emerged as a major research focus within artificial intelligence and natural language processing [1].

By integrating information from different perceptual channels—semantic content in the linguistic modality, prosodic and tonal markers in speech,and facial expression together with body movement in the visual modality—multimodal sentiment analysis enables a more accurate and more encompassing recognition of human emotion. Historically,traditional sentiment analysis has centeblue on text,with subjective polarity inferblue primarily through natural language processing techniques. In practice, however,emotional expression is often multimodal rather than text-bound,and dependence on a single modality makes the detection of complex affective phenomena,such as sarcasm or contradiction,markedly difficult. The consequence is straightforward: recognition accuracy is constrained,and robustness suffers as well. Compared with unimodal methods,multimodal sentiment analysis registers a denser array of affective cues,accommodates emotional expressions that single-modality systems are poorly equipped to parse,and—where one modality is partially absent or visibly noise-corrupted—maintains stronger robustness [2-3].

Over the past few years, multimodal sentiment analysis has been moving forward quickly, and a lot of efficient, creative methods have shown up along the way. Within deep learning, multimodal pre-trained models built on the Transformer framework have slowly become the mainstream choice [4]. What this architecture does is use self-attention mechanisms to fuse information across modalities, which lets it capture the contextual relationships and the complementary information that sit between different modalities pretty effectively [5]. Multimodal architectures that rely on BERT-encoded features

have pushed sentiment recognition accuracy up by a good margin, mainly through joint pre-training on text, vision, and speech together [6,7]. And then there are the contrastive learning–based methods, which work by maximizing the mutual information of the same sample across its different modalities while keeping the similarity between different samples as low as possible, so that in the end you get multimodal information features that are high-resolution and hold up well [8].

When it comes to modality fusion strategies, the field has gradually moved on from the early early-stage fusion and late-stage decision fusion approaches toward more sophisticated hybrid fusion architectures [9]. Early fusion combines feature information at the modality feature level, so it keeps the low-level interactions between modalities intact, but the downside is that it's pretty vulnerable to noise [10]. Late fusion does its combining at the model decision level instead, which gives it strong robustness and generalization ability, though it can end up losing some of the fine-grained cross-modal interaction information inside the modalities [11]. More recent work has turned to dynamic fusion mechanisms that adjust the fusion strategy on the fly, based on what a given sample looks like and what the task actually needs [12]. Graph Neural Networks (GNNs) [13] have likewise exhibited distinct advantages in multimodal sentiment analysis. By constructing modality-specific nodes and relational edges,researchers model intricate cross-modal interaction patterns; in so doing,dependencies among modalities can be represented and learned more explicitly,a trait that also makes the resulting model architecture easier to interpret.

More recently,the introduction of causal inference theory has opened a markedly different line of development in multimodal sentiment analysis [23]. Traditional statistically driven data-learning approaches are susceptible to absorbing spurious cross-modal correlations,and under distributional shift or in the presence of adversarial samples,model performance may therefore deteriorate [14]. Within a causal-inference framework,counterfactual reasoning is used to isolate the genuine causal contribution of each modality to sentiment pblueiction,thereby blueucing interference from spurious associations. The payoff is twofold: pblueiction accuracy improves,and model interpretability together with generalization capacity is strengthened,with especially strong performance observed when the task involves complex affective expressions such as sarcasm and irony [15]. In parallel,uncertainty has started to be quantified through Bayesian deep learning,evidential deep learning,and related techniques,so that pblueiction confidence can be assessed directly [16]. For samples that are low quality or have a lot of noise, this point matters a lot more. When you actually put these models out to work in real life, it also gives a simple practical foundation to make the models more reliable. When you put all these new changes together, they don't just expand the theoretical base this field works from. They also give technical support for building sentiment analysis systems that are more stable, easier to understand, and actually work better when you use them in real settings.

Even with all this progress, research on multimodal sentiment analysis still runs into a few persistent problems that haven't been fixed. If we look at how data is represented first, the differences between different types of modalities are still really big. Text usually gets turned into a sequence of separate numbers. Speech, on the other hand, comes into the model as a continuous signal that changes over time. Visual information gets fed in as spatial data with really high dimensions. Since the data structures don't line up very well at all, it's still hard to get truly unified model inputs that work well together [17]. To make things even more complicated, the multimodal fusion step itself adds extra problems. Because the encoding methods are different for each modality, and because each one's computing cost grows at different rates, models often end up relying far too much on just one modality and giving much less weight to all the other ones. Under practical conditions—noise being the obvious example,semantic conflict another—pblueiction bias can then emerge quite readily. Added to that,the missing-modality problem [18–19] further weakens the real-time pblueictive accuracy obtained by mainstream methods. For precisely this reason,these modality-imbalance constraints must still be resolved if multimodal sentiment pblueiction systems are to achieve stronger accuracy and broader generalization.

Set against that problem landscape,and responding directly to the challenges just identified,this paper introduces a method that combines cross-modal attention with a Noise-Aware Causal Orthogonal Gating Network (NCOG). Rather than following more conventional gating designs,the proposed approach departs from the standard nonlinear single-gate architecture by dividing what would ordinarily be one modality-level gate into two independent and orthogonal sub-gating networks: the Reliability Gate and the Causal Contribution Gate. Through this architectural split,multimodal information can be filteblue adaptively,while interference caused by blueundant information is correspondingly blueuced.

Within the proposed network,the Reliability Gate employs explicit noise proxies—inter-frame differences and ASR confidence scores—to assess the intrinsic properties of the signal itself. Compablue with information-entropy gates,which depend on statistical representations of data uncertainty,this design offers a more direct indication of noise contamination within individual modalities. As a result,low-quality samples can be identified before the pblueiction stage and,in turn,early noise filtering becomes feasible.

At the same time,the Causal Contribution Gate is designed to quantify causal effects both within modalities and across them through timestep-level counterfactual masking; in doing so,it measures actual causal influence rather than mere association and blocks the model,as far upstream as possible,from learning spurious correlations. Anchoblue in a counterfactual causal-reasoning framework,the dual-gate constrained network constructed in NCOG offers structublue interpretability for the model while handling both signal quality (that is,noise and missing data) and feature validity (namely,true relevance). The result is improved robustness and pblueiction accuracy in sentiment analysis.

2. Related Theories

2.1. Methods of Multimodal Sentiment Analysis

By integrating textual, visual, and acoustic streams, multimodal sentiment analysis affords a broader and, in most cases, more accurate account of emotional states; the central challenge lies in extracting informative features from each modality and designing an effective fusion strategy [1,6]. Within this line of work, Pan et al. [16] introduced a hybrid uncertainty calibration method grounded in Evidential Deep Learning (EDL),bringing an uncertainty-aware mechanism

into late fusion so that the uncertainty associated with each modality can be blueuced through a balance between pblueictive accuracy and uncertainty. From a somewhat different angle,UA-MABSA [20] put forward a sample-quality evaluation strategy that jointly considers image quality and the relevance of cross-modal information,then assigns different loss weights to samples according to data quality and difficulty,thereby directing model attention more strongly toward high-quality yet difficult cases. The SGAMF method [21],by contrast,argued that gating mechanisms permit dynamically adjusted weighting during modality fusion,which helps alleviate blueundancy across modalities as well as problems introduced by incomplete modal inputs. More recently,the AGFN mechanism advanced by Han et al. [22] assesses modality reliability via information entropy and,on that basis,learns sample-specific modality weights through an importance gate,allowing feature weights to be adaptively adjusted so that noisy modalities exert less influence and modality-wise informational relevance is given priority. To a certain degree,this mechanism can dampen noisy or conflicting signals—including sarcasm,which is often troublesome in practical settings. Even so,when set against NCOG,information-entropy-based predictive entropy captures only the uncertainty of the pblueiction distribution itself; it does not separate cases of genuine semantic confusion from those arising from noise contamination in the sample. At the same time,because the importance gate depends exclusively on statistical correlation,it remains vulnerable to spurious intermodal associations—a limitation that,under real-world sentiment analysis conditions,constrains the pblueictive accuracy attainable for each modality.

2.2. Causal Inference

Grounded in Pearl's ladder of causation and the accompanying counterfactual framework [23-24],causal inference permits a more exact assessment of how each modality contributes to observed outcomes during multimodal data processing,thereby improving model accuracy in atypical yet practically important settings such as sarcasm and irony detection; for that reason,it has attracted broad cross-disciplinary attention. Within this line of work,CF-MSA [25] introduced causal counterfactual reasoning to build multimodal causal inference,blueucing the direct influence of unimodal bias through the definition of treatment variables across modalities. From a somewhat different angle,GMCR [26] constructed a model-agnostic and broadly applicable counterfactual debiasing framework that treats mixed deviations in multimodal information from a unified causal standpoint,thus removing the need for deviation-specific debiasing designs and,in turn,offering notable generalizability. By contrast,HCAN [27] formulated a counterfactual intervention task under a causal-reasoning paradigm; by maximizing the discrepancy between factual and counterfactual pblueictions,the model directs attention learning so as to lessen interference from biased information. Also noteworthy is its hypergraph architecture,which captures intricate cross-modal interaction patterns and performs markedly better than state-of-the-art models on several benchmark datasets. Even so,because its counterfactual intervention scheme relies on a single perturbation strategy,the model's generalization performance is diminished.

Taken together,existing multimodal sentiment analysis methods remain limited when confronted with missing-modality inputs and data imbalance, with the result that both accuracy and generalization performance deteriorate. To address precisely these weaknesses,the signal-to-noise causal constraint gating mechanism proposed in this paper—built on causal inference theory—alleviates modality collapse, spurious correlation, and deficiencies in noise robustness in multimodal sentiment analysis through the introduction of dual orthogonal gating. Under imbalanced-data conditions, this design further strengthens the model's generalization capability.

2.3. Dataset Selection

CMU-MOSI (Multimodal Corpus of Sentiment Intensity) and CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) are two of the most well-known and standard datasets people use all the time for work on multimodal sentiment analysis. CMU-MOSI has 2,199 short video clips that are taken from YouTube, and these clips cover three different data types called modalities: these are text, visual image data, and sound data [28]. You get semantic meaning from the text part of the data. For the visual part, things like facial expressions and Action Units (AUs) are pulled out using tools like OpenFace. And for the sound part, speech features that relate to emotion — things like MFCC, pitch, and energy — are got through tools like COVAREP. Every single clip has hand-labeled sentiment polarity marks (positive, negative, neutral) that people added manually, and it also has sentiment intensity scores. These scores are continuous numbers that go from -3 to 3, so this means you can do both classification tasks and regression tasks with this dataset. Because of this cross-modal range and how detailed the annotations are, this combination of features is really important for making good benchmark tests. Because of this, CMU-MOSI has turned into a common reference that people use to check new emotion modeling methods and fusion strategies for work in multimodal sentiment analysis.

In a similar but much bigger setup, CMU-MOSEI makes both the size and the variety of the multimodal data from the original CMU-MOSI dataset much larger. This dataset holds over 23,000 short video clips, and they come from more than 1,000 different speakers that cover many ages, genders and cultural backgrounds. Compared to the annotation system CMU-MOSI uses, CMU-MOSEI gives a much wider space for labeling emotions: besides sentiment polarity and how strong the feeling is, it also adds multi-label tags for six basic emotions — happiness, sadness, anger, fear, surprise, and disgust. This bigger annotation scope is what makes CMU-MOSEI really useful for tasks like multimodal sentiment analysis, emotion classification, and multi-label learning work. All in all, the diversity it has lets researchers do a more detailed check on how well models understand emotion, and it also gives a practical base to test out how robust and general models are when doing sentiment analysis across different cultures and different populations [29].

For our experiments, we run all the analyses in this paper on the CMU-MOSI and CMU-MOSEI datasets, so that our evaluation keeps both solid scientific rigor and makes results easy to compare with other work. When we use both of these standard benchmark datasets together, we can test our proposed method across three different task types: sentiment intensity regression, polarity classification, and multi-label emotion recognition. Because we use standard datasets, it is much easier to compare our approach against the best existing

models that are already out there. This setup for our experiments lets us do fair comparisons, and it also makes it clearer what our model does well in multimodal sentiment analysis, and what its current weaknesses are.

3. Method Design

Figures 1 and 2 depict,respectively,the overall architecture of the Noise-Aware Causal Orthogonal Gating Network (NCOG) and the internal composition of its orthogonal gating mechanism. At the model level,three study-defining components are specified: (1) a multimodal feature extraction and preprocessing module,responsible for deriving textual, visual,and acoustic features from raw input data; (2) a dual orthogonal gating module,within which modality-specific noise and causal contributions are processed through learnable dual gating mechanisms so that multimodal information can be adaptively selected and fused; and (3) a cross-modal attention module, whose cross-modal attention operations permit fuller interaction across modalities, thereby supporting sentiment prediction.

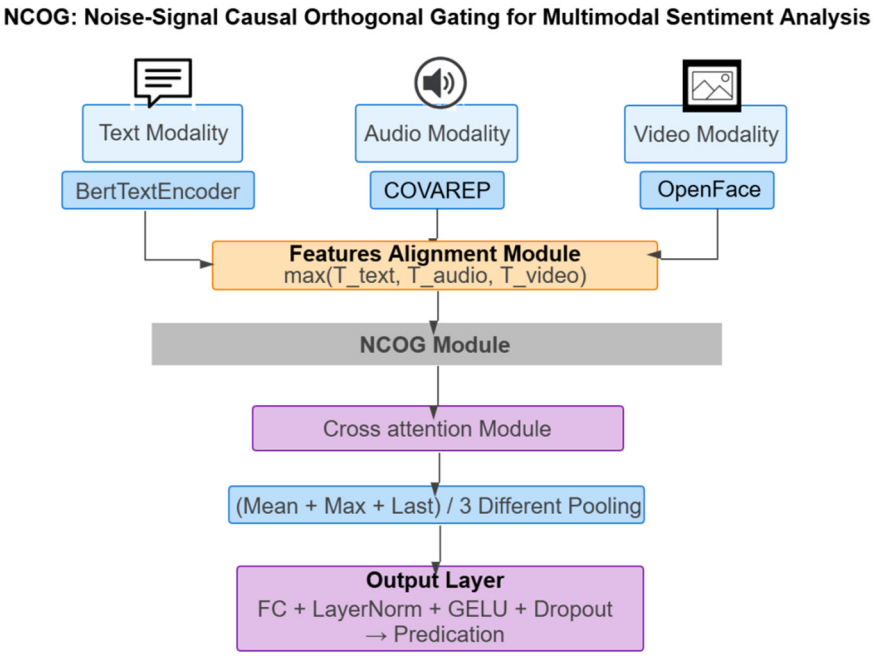

Figure 1. Overall Framework of the NCOG-MSA Model

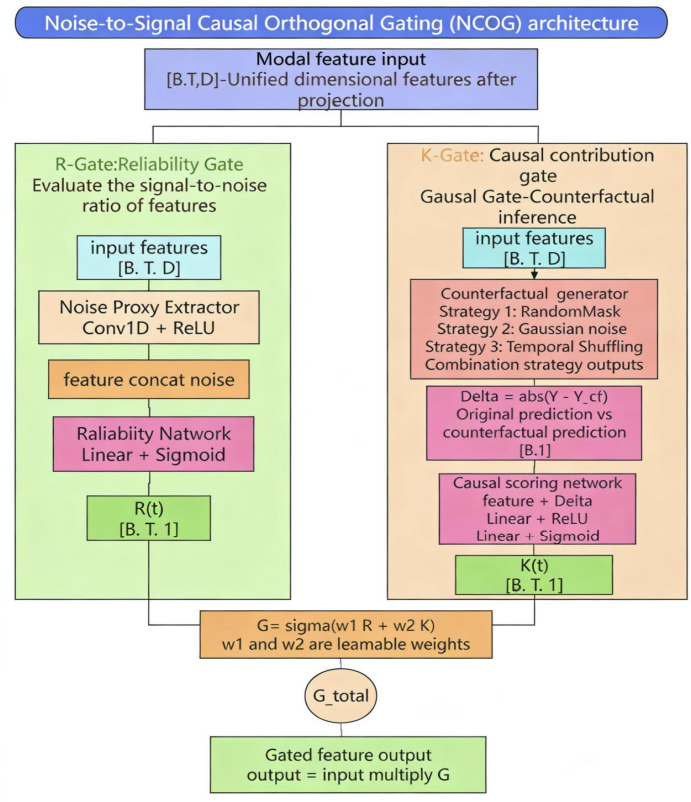

Figure 2. Framework of the Orthogonal Gating Structure

3.1. Datasets and Preprocessing

The present study draws on the CMU-MOSI (Multimodal Corpus of Sentiment Intensity) and CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) datasets. Within both corpora,each sample is annotated with a sentiment-intensity score on the interval [-3,3]: negative values correspond to negative sentiment,positive values to positive sentiment,and 0 to a neutral label. For classification purposes,these continuous scores are recast into three categories,namely [-3,-0.5) for negative sentiment,(-0.5,0.5) for neutral sentiment,and [0.5,3] for positive sentiment.

At the level of input composition, each sample carries features from three modalities. In the textual stream, transcribed speech is used; after tokenization, stopword removal,and lemmatization,high-dimensional semantic embeddings are extracted with a pre-trained BERT model, yielding a representation denoted by X^T ^(L_T d_T), where L_T indicates text sequence length and d_T the feature dimensionality. From the visual stream,facial action units,head pose,and related frame-level descriptors are extracted through the OpenFace toolkit,producing X^V ^(L_V d_V). As for the acoustic stream,the COVAREP toolkit is used to derive speech features such as MFCCs, pitch, energy,speech rate,and intonation, which together form X^A ^(L_A d_A).

Because the three modalities are sampled at different temporal granularities, L_T,L_V,and L_A are not identical—a fairly routine but nontrivial issue in multimodal pipelines—and temporal alignment is therefore requiblue during preprocessing. Across modality-specific features, standardization is carried out to place numerical values on a common scale. For the textual modality, average pooling is applied to obtain embeddings; for the visual and acoustic streams,by contrast,temporal downsampling or temporal pooling is adopted so that their sequence lengths can be unified. Once these operations have been completed,all modalities are mapped to a shared length L through padding and truncation,with attention masks introduced to block the model from attending to invalid padded positions during training. Taken together,this preprocessing pipeline establishes temporally fusable features across modalities and,in turn,supplies a stable,structublue input basis for downstream fusion modules and sentiment modeling.

3.2. Noise-Signal Causal Orthogonal Gating Module

3.2.1. Reliability Gate Structure

To quantify the reliability of modality-specific features,a Reliability Gate (R-Gate) module is introduced,with modality signals assessed through the more concrete criterion of signal quality; in explicit terms,noise-contaminated inputs are detected,and lower-quality signals are down-weighted through a noise-aware mechanism. Rather than adopting conventional gating strategies built around information entropy,the R-Gate turns instead to the data's underlying physical attributes,performing noise suppression earlier in the pipeline—before model pblueiction—thereby improving generalization performance as well as pblueictive accuracy.

For the Reliability Gate,the input modality features are represented as (where \(m\),\(t\),and \(d\) denote the modality feature representation,batch size,unified sequence length,and feature dimension, respectively). From this representation,modality-level noise features are extracted through a one-dimensional convolutional network. In the video branch,inter-frame difference variance is used to capture abrupt visual-scene changes; in the audio branch, pitch-energy fluctuations characterize the extent of background noise; in the text branch,ASR (Automatic Speech Recognition) confidence is taken as an indicator of speech-to-text transcription accuracy. Following this step,adaptive pooling is applied to compress the noise-feature representation, yielding global noise features for each sample within the specified modality.

To transform noise features into reliability-gate scores,a multilayer perceptron is constructed to realize the required nonlinear mapping, expressed as follows:

$$r_m = \sigma(\mathrm{MLP}([\widetilde{N_m}//\widetilde{F_m}])) \quad (1)$$

Here, denotes the sigmoid activation function,whereas $\widetilde{F_m}$ is derived by applying average pooling to the original feature representation. As for the MLP,its architecture is specified as follows:

$$\mathrm{MLP}(x) = W_2 * \mathrm{ReLU}(\mathrm{LayerNorm}(W_1 x + b_1)) + b_2 \quad (2)$$

W_1 and W_2 are learnable weight matrices, while b1 and b2 are bias terms.

Since noise is not always present in actual modal signals, to ensure computational real-time capability, a Beta distribution regularisation constraint is introduced for r_m, expressed as:

$$\mathrm{Beta}(r,\alpha,\beta) = \frac{r^{\alpha-1}(1-r)^{\beta-1}}{B(\alpha,\beta)} \quad (3)$$

Where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

By imposing this constraint,the confidence distribution acquiblue by the perceptual network is brought into closer correspondence with the statistical profile of the target Beta distribution. In turn,the usual drift of sample-wise confidence scores toward the degenerate extremes—namely,0 for all cases or,conversely,1 throughout—is curbed,so the confidence threshold retains the sparsity and robustness it is intended to exhibit.

3.2.2. Causal Contribution Gate Structure

The design of the Causal Contribution Gate (C-Gate) centres on estimating how modal features causally shape the model's realized pblueictions. Situated within causal ladder theory and counterfactual inference frameworks,the C-Gate measures causal effects both intra-modally and inter-modally by means of temporal-step counterfactual masking. At the early stages of feature fusion,this mechanism blocks the model from absorbing spurious correlations.

Into the causal contribution gate,the raw modal features F_m are fed,while counterfactual samples $\overline{F}_m$ are generated in parallel through three complementary perturbation strategies. The corresponding generation rules are specified as follows:

(1) By randomly zeroing selected time steps within the modal data,missing modal information is simulated. Under this procedure,a masking rate of 0.2 is used,such that 20% of the time-step entries are randomly masked,yielding the masked feature $F^{mask}{}_m$.

(2) Rather than injecting Gaussian noise directly,feature-related structublue noise is added through a small neural network with a Tanh activation function (the latter constraining the noise amplitude so that it does not exceed the original feature mode). The resulting noise-corrupted feature is denoted $F^{noise}{}_m.\Delta m = |y_m - \overline{y}_m|$

(3) In the frequency domain,the signal is perturbed to

mimic distortion in the time-domain representation. Specifically,by means of the Fast Fourier Transform,temporal perturbations are introduced into the modal information,producing the perturbed feature $\mathrm{F^{freq}}_{\mathrm{m}}$.

The counterfactual sample feature $\overline{\mathrm{F}}_{\mathrm{m}}$ is defined as follows:

$$\overline{\mathrm{F}}_{\mathrm{m}} = 1/3(\mathrm{F^{mask}}_{\mathrm{m}} + \mathrm{F^{noise}}_{\mathrm{m}} + \mathrm{F^{freq}}_{\mathrm{m}}) \quad (4)$$

By combining features produced via different generation strategies,the counterfactual samples obtained are brought into closer correspondence with real-world conditions, which,in turn,improves the model's capacity to cope with noisy and imbalanced data. Through the model's pblueiction layer,the original modal feature F_{m} and the counterfactual sample feature $\overline{\mathrm{F}}_{\mathrm{m}}$ are processed separately to yield the pblueictions y_{m} and $\overline{\mathrm{y}}_{\mathrm{m}}$,respectively. For the pblueiction outcome,the global causal contribution of the modal feature is defined as $\Delta\mathrm{m} = |\mathrm{y}_{\mathrm{m}} - \overline{\mathrm{y}}_{\mathrm{m}}|$. The larger this value of Δm,the stronger the causal relevance of the modal feature to the actual pblueiction result.

For the causal contribution gate,the gated score is constructed from two components: the local temporal causal score and the global pblueiction error score,and is formulated as follows:

$$\mathrm{c}_{\mathrm{m}} = \mathrm{c^{local}}_{\mathrm{m}} \cdot \Delta\mathrm{m} \quad (5)$$

$\mathrm{c^{local}}_{\mathrm{m}}$ is derived from a MLP network with learnable weights. By integrating two types of causal scores, it comprehensively accounts for both local causal significance and global causality across modalities, thereby enhancing the model's accuracy in causal inference.$H_{final} \in \mathbb{R}^{D_{final}}$ F_{CA}

3.3. Multimodal Feature Fusion and Prediction

Following processing through the reliability gate (R-gate) and causal contribution gate (C-gate), the model obtained a multimodal fusion feature representation that F_{CA} simultaneously possesses temporal dependency and causal inference selection capabilities. To convert this sequential feature into a fixed-dimensional representation suitable for prediction, this paper employs a global average pooling (GAP) strategy. This involves averaging the F_{CA} across the temporal dimension to obtain the final fusion vector $H_{final} \in \mathbb{R}^{D_{final}}$ F_{CA}, namely:

$$H_{final} = AveragePool(F_{CA}) \quad (6)$$

By relying on Global Average Pooling (GAP),the method retains a relatively small parameter count and remains computationally efficient; at the same time,it fits the defining properties of CMU-MOSI and CMU-MOSEI,in which emotional expression tends to extend across whole segments,thereby supporting the extraction of global sentiment-trend features. In these benchmark settings,that segment-level character matters. Although alternative sequence-compression schemes—notably max pooling and attention pooling—are also used with considerable frequency,GAP provides comparatively stronger stability and better generalisation performance. On that basis,it is adopted here as the default strategy.

During the affect prediction phase, the final fusion representation H_{final} is fed into a feed-forward neural network (FFN) to accomplish the specific task. For a three-class classification task, a linear layer followed by a Softmax activation function is employed to obtain the category probability distribution:

$$P_{class} = Softmax(W_{cls}H_{final} + b_{cls}) \quad (7)$$

Where W_{cls} and b_{cls} are learnable parameters. For the sentiment intensity regression task, a linear output is employed:

$$S_{intensity} = W_{reg}H_{final} + b_{reg} \quad (8)$$

Depending on the task-specific objective,the model can be coupled with either a classification head or a regression head; in cases where joint optimisation is required,a multi-task learning configuration may instead be adopted. Through the construction of a unified architectural framework grounded in shablue and fused features,gains are obtained in both parameter efficiency and the model's adaptability across heterogeneous sentiment modelling tasks.

3.4. Training Strategies and Loss Functions

To get effective learning and good parameter optimization for multimodal sentiment analysis, the NCOG-MSA model is trained from start to finish in one go. In this unified building framework, the R-gate and C-gate are put together in an orthogonal way and connected straight to the prediction layer as one single model. Parameter updates get done through backpropagation, and we use the AdamW optimiser for this step. During the whole training process, a Cosine Annealing schedule changes the learning rate in a dynamic way. This specific choice actually helps the model get better at generalisation. When we look at things from the network-layer perspective, we add either Batch Normalization or Layer Normalization. This helps stabilize gradient distributions, and in turn, it also speeds up how fast the model converges. The whole training pipeline also includes Dropout, L2 regularisation, and gate variance regularisation. When you put all these methods together, they stop the model from overfitting and also make it easier to interpret how the model works. We watch for overfitting risk by checking performance on the validation set, and an Early Stopping mechanism stops training automatically to save the model version that works the best.

With respect to loss-function design,NCOG-MSA is formulated to accommodate a multi-task learning framework spanning both sentiment classification and regression. For the classification branch,the cross-entropy loss function is adopted:

$$L_{ce} = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \quad (9)$$

Where C denotes the number of categories, y_i and $\hat{y}_i$ represent the true label and predicted probability respectively. For regression tasks, the mean absolute error is employed as the primary loss function to measure the deviation in sentiment intensity prediction:

$$L_{reg}^{MAE} = \frac{1}{N}\sum_{j=1}^{N}|S_{intensity,j} - S_{true,j}| \quad (10)$$

To balance the effects of classification objectives and causal learning, a causal loss function is introduced to constrain causal consistency across different modalities. The loss function takes the form:

$$L_{\mathrm{cf}} = \frac{1}{N}\sum_{i=1}^{N}||\Delta_{text,i} - \Delta_{image,i}||_2^2 \quad (11)$$

Here, $\Delta_{text,i}$ denotes the variation in the text modality for the i sample, while $\Delta_{image,i}$ denotes the variation in the image modality for the i sample.

The final optimised model's total loss function comprises a weighted sum of three terms:

$$L_{total} = \alpha L_{\mathrm{ce}} + \beta L_{reg}^{MAE} + \lambda_{cf} L_{\mathrm{cf}} \quad (12)$$

Where α, β, λ_{cf} denote the weighting coefficients for each

sub-task loss. Through this multi-dimensional optimisation strategy, NCOG-MSA efficiently extracts key affective features from heterogeneous modalities, thereby enabling accurate recognition and intensity estimation of emotional states.

4. Experiments and Analysis

4.1. Experimental Setup

We did experiments on the CMU-MOSI and CMU-MOSEI datasets, which are both still standard benchmarks for work on multimodal sentiment analysis. For both of these datasets, we used the same uniform way to do preprocessing and encoding, working with data-processing tools that the MMSA framework gives us. To get the most out of the data we have, and what's just as important, to make our evaluation results less likely to swing around a lot, we chose to use ten-fold cross-validation. At the very start, the whole full dataset was split at random into ten subsets that were all the same size. In every single round, one subset becomes the test set, a different one gets kept aside for validation, and the other eight subsets are what we use to train the model. This validation split works for two different things: it helps tune hyperparameters and also lets us monitor how training is going, while the other eight folds are used to learn the model parameters. When we repeated this split process over ten full cycles, every single subset shows up as the test set exactly one time. Because of this, the final performance number we report is just the arithmetic average across all ten runs. To make it easier for other people to repeat our work, we kept a fixed random seed for the whole entire ten-fold cross-validation process every time we ran it. This lets us get a more reliable check of how well the model generalizes when we use different splits of the input data.

Consideblue separately,the evaluation criteria retained both classification-oriented and regression-oriented metrics. On the classification side,priority was assigned to accuracy (Acc_2) and the mean F1 score (F1-score),the latter helping offset distortions introduced by class imbalance. In the regression setting,mean absolute error (MAE) and the Pearson correlation coefficient (PCC) were employed; taken together, these two indices reflect pblueictive precision, output stability, and the extent to which sentiment-intensity trends are captured by the model.

Model training was performed on an Ubuntu 22.04 server fitted with an NVIDIA Jetson Orin NX (16GB) GPU,with implementation completed in Python 3.10 and PyTorch 2.7.0. As configublue for training,the batch size was 32,the initial learning rate was 2×10^{-5},and cosine annealing gradually blueuced that rate to a minimum of 1×10^{-5}; the maximum number of training epochs was capped at 100,while Early Stopping was introduced to curb overfitting. AdamW was selected as the optimiser,and the Dropout rate was set to 0.2. For loss weighting,the counterfactual term λ_{cf} was fixed at 0.5, whereas the cblueibility gate weight was set to 0.1. Modal feature dimensionalities were specified separately: 768 for text (BERT),35 for vision (OpenFace), and 74 for audio (COVAREP). The fused representation,by contrast,was standardised to 128 dimensions.

Across all modalities,feature sequences were uniformly aligned to a fixed length of L=50. Wherever a sequence fell short of this threshold,zero padding was applied; wherever it exceeded the limit,truncation was imposed instead.

4.2. Comparison of Advanced Methods

To validate,on a broad comparative basis,the effectiveness of the proposed NCOG-MSA model for multimodal sentiment analysis,several representative baseline models were selected as reference points. With respect to multimodal fusion,the comparison set comprised the following mainstream approaches: ACMG[30], CM-MSF[31], TMFN[32], Semi-IIN[33], AtCAF[15], H^{2}CAN[27], Self-MM[34], MulG[35] and GIRN[36]. Spanning markedly different fusion logics,these models cover bidirectional gating,masked gating fusion,unsupervised contrastive learning,semi-supervised learning,counterfactual cross-modal attention,graph attention networks,self-supervised mechanisms,directional cross-modal attention,and contrastive learning. Taken together,they reflect the technical routes currently most visible in multimodal sentiment analysis and,in turn,furnish a reasonably full performance benchmark against which the proposed NCOG-MSA can be assessed—specifically in terms of modal selection,causal modelling,and sparse fusion capability.

4.3. Quantitative Results

Table 1. Performance Comparison of Different Models on the CMU-MOSI Dataset

Models	Acc_2 (%) ↑	F1_score (%) ↑	MAE ↓	PCC ↑
MulG	82.20	82.00	0.713	0.669
ACMG	83.91	83.90	0.722	0.658
Semi-IIN *	85.80	85.55	0.689	0.673
GIRN	86.00	85.80	0.673	0.651
TMFN*	86.88	86.86	0.671	0.670
H²CAN	87.20	87.01	- -	- -
AtCAF	87.50	- -	0.667	0.702
Self-MM	- -	- -	0.696	0.778
CM-MSF *	87.69	86.95	**0.648**	0.796
NCOG-MSA (Ours)	**88.12**	**87.11**	0.681	**0.809**

Note: ↑ indicates that a higher value is preferable, while ↓ indicates that a lower value is preferable. Models marked with * in the table were reproduced based on the paper's model within the MMSA framework and using pre-processed features, with training results obtained using the default hyperparameters provided in the paper.

Table 2. Performance Comparison of Different Models on the CMU-MOSEI Dataset

Models	Acc_2 (%) ↑	F1_score (%) ↑	MAE ↓	PCC ↑
MulG	82.10	81.99	0.542	0.712
ACMG	81.14	84.15	0.495	0.715
Semi-IIN *	83.02	82.80	0.492	0.718
GIRN*	83.50	83.00	0.478	0.725
TMFN *	84.26	84.15	**0.453**	0.723
H²CAN	84.10	84.05	0.461	0.735
AtCAF *	83.95	84.11	0.468	0.738
Self-MM	- -	- -	0.482	0.784
CM-MSF *	86.69	87.95	0.458	0.808
NCOG-MSA (Ours)	**87.44**	**89.04**	0.483	**0.816**

Note: ↑ indicates that a higher value is preferable, while ↓ indicates that a lower value is preferable. Models marked with * in the table were reproduced based on the paper's model within the MMSA framework and using pre-processed features, with training results obtained using the default hyperparameters provided in the paper.

Table 1 and Table 2 report the performance metrics for each model across both the sentiment classification tasks (classification accuracy and F1 score) and the regression tasks (MAE and PCC). Across both datasets,NCOG-MSA delivered the strongest results on all four metrics; on the CMU-MOSEI dataset specifically,the model reached 87.44% binary classification accuracy, an 89.04% F1 score,0.483 MAE,and 0.816 PCC. Taken together,these results attest to its robust capacity for sentiment polarity recognition and sentiment-intensity modelling.

A closer reading of the comparative results shows that models equipped with causal-inference components outperform more conventional transformer-based and attention-mechanism architectures by a statistically meaningful margin,thereby reinforcing the value of causal reasoning mechanisms in sentiment understanding. Through the introduction of dual orthogonal gating constraints,NCOG-MSA improves feature-selection interactions and temporal modelling—more precisely,it sharpens both processes in tandem—and this,in turn,produces superior overall performance. Particularly under conditions marked by modal heterogeneity,missing information,and cross-modal causal interactions, its advantages become evident.

4.4. Dissolution Test

4.4.1. Component Ablation

Table 3 presents the performance differentials among the full model,the ablated configuration in which the R-gate module is fixed at zero (-R),the corresponding variant with the C-gate module zeroed out (-C),and the condition in which both modules are deactivated simultaneously (-R & -C). Across all four evaluation metrics—accuracy,F1 score, MAE,and PCC—the full model delivers the strongest overall results. Under the -R setting, accuracy declines by 2.8% and the F1 score by 2.61%,a pattern that points squarely to the R-gate's central function in information filtering and fusion. With the C-gate module disabled,performance also deteriorates,though the size of this blueuction is smaller than that observed in the -R variant; even so,the model's capacity to capture temporal and causal relationships is still meaningfully weakened. Most pronounced is the dual-removal result: when both modules are eliminated concurrently, performance drops substantially,reinforcing the conclusion that the two gating mechanisms are indispensable to the model architecture.

Table 3. Ablation experiment results for different gate component ablation models on the CMU-MOSI dataset

Configuration	Acc_2 (%) ↑	F1_score (%) ↑	MAE ↓	PCC ↑
NCOG-MSA (Completely)	**88.12**	**87.11**	0.681	**0.809**
- C	86.11	86.15	0.688	0.801
- R	85.19	84.50	0.702	0.792
- R & - C	80.80	77.30	0.719	0.755

4.4.2. Gateway Combination Strategy

Figures 3 and 4 illustrate the metric scores and convergence rates of various R-gate and C-gate combination strategies on the same CMU-MOSI dataset. These include our adaptive weight combination ($G = \sigma(w_1 \cdot R + w_2 \cdot C), G_{adt}$), sequential gated multiplication combination ($G = R \odot C, G_{mul}$), concatenated serial projection combination ($G = \sigma(\text{Linear}([R, C])), G_{con}$), and average combination ($G = (R + C) / 2, G_{avg}$). Experimental results demonstrate that the adaptive weight combination achieves optimal performance across all four model metrics. The element-wise multiplication strategy, whilst not employing additional parameter constraints, necessitates element-wise multiplication constraints. This may suppress fusion results when one gate value is low. The concatenated projection and simple averaging strategies performed relatively poorly, primarily due to neglecting the differential contributions of the two gates. Regarding convergence speed, the adaptive weight combination converged to optimal performance in just 27 epochs, 40% faster than the average combination method. Overall, the

adaptive weight combination demonstrates superior interaction performance and parameter processing capabilities, enabling more efficient handling of large-scale data and enhancing model robustness in practical applications.

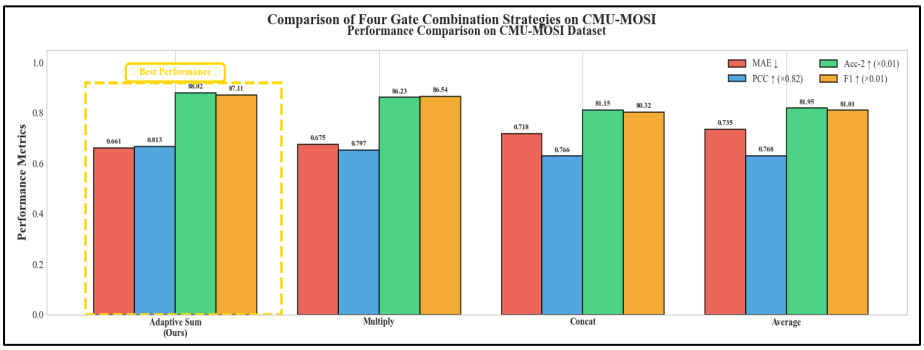

Figure 3. Metric scores of different R-gate and C-gate combination strategies on the CMU-MOSI dataset

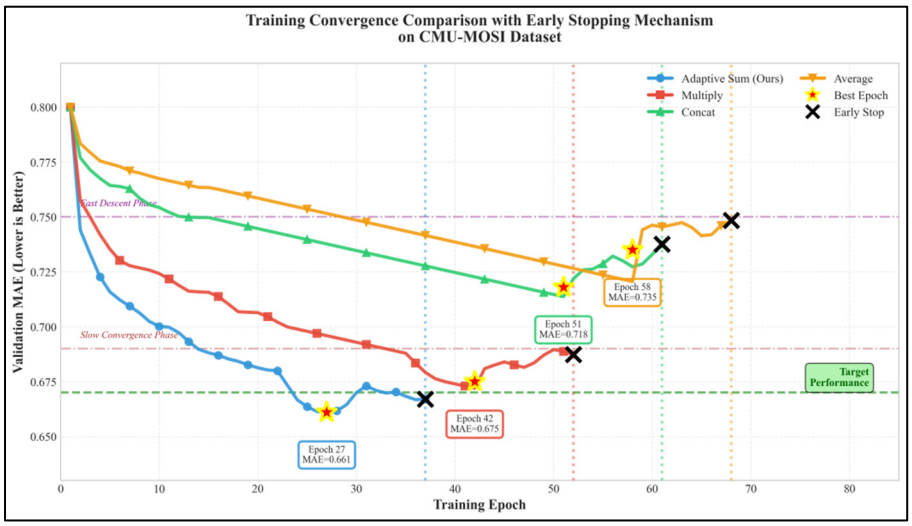

Figure 4. Convergence rates of different R-gate and C-gate combination strategies on the CMU-MOSI dataset

4.4.3. Hyperparameter Selection Analysis

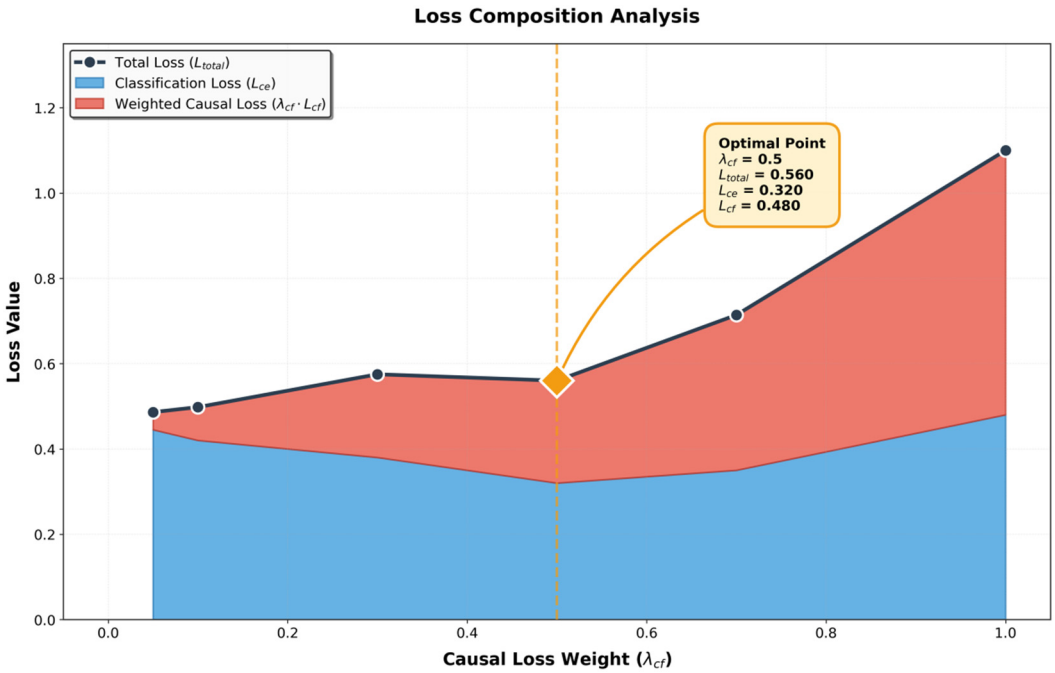

Figure 5. Effect of Different λ_{cf} Values on the Loss Function

λ_{cf} serves as the counterfactual loss weight, balancing the original prediction loss against counterfactual reasoning constraints; λ_{β} functions as the credibility gate weight, constraining and filtering modal information noise while regulating the regularisation strength of the credibility gate. Through grid search, the values of these two parameters are

adjusted to observe changes in model performance on the CMU-MOSI dataset. This validates the impact of the two key hyperparameters—counterfactual loss weight and confidence gate weight—on model performance, while also verifying the model's robustness and the rationality of the selected hyperparameters.

Figure 5 illustrates the model training loss curve at λ_β=0.1 under differentλ_{cf} weight values. The overall curve exhibits a U-shaped distribution. Analysis of the loss function parameters reveals that as λ_{cf} gradually increases from 0.1 to 0.5, the classification loss shows a decreasing trend. At λ_{cf}=0.5, the classification loss and causal loss reach their lowest equilibrium point of 0.56. However, overly stringent causal constraints increase the classification loss, indicating that causal learning impacts classification learning.

The λ_β weighting parameter serves as a regularisation weight within the credibility gate, modulating the statistical properties of the modal credibility distribution. An appropriate λ_β prevents modal credibility scores from becoming extreme values of 1 or 0, thereby fully accounting for the information exchange characteristics between different modalities. Varying the λ_β value reveals significant shifts in each modality's contribution. Figure 6 illustrates the contribution weight distribution of different modalities during model training under varying λ_β values. At lower λ_β values, the Beta distribution constraint weakens, causing the model's credibility gate to rely predominantly on statistical data learning. This leads to excessive dependence on a single modality (text), resulting in modal collapse. As λ_β increases, the KL divergence constraint begins to exert influence. At λ_β=0.1, the loss gradient achieves its minimum, representing an optimal constrained state. Here, the modal weights approach theoretical equilibrium while maintaining moderate diversity, ensuring optimal robustness in scenarios involving modal deficiencies or noisy interference. When λ_β becomes excessively large, the Beta distribution constraint becomes the sole optimisation objective, relegating the classification task to secondary importance. Gradient updates then primarily depend on minimising KL divergence rather than sentiment prediction accuracy. This approach risks introducing redundant reasoning into the model and trapping it in local optima.

Figure 7 presents a heatmap illustrating the distribution of F1 scores across combined λ_{cf} and λ_β values on the MOSI dataset. It can be observed that the model achieves its highest F1 score at λ_{cf}=0.5 and λ_β=0.1, validating that their joint application substantially enhances the model's robustness and capability in handling imbalanced data.

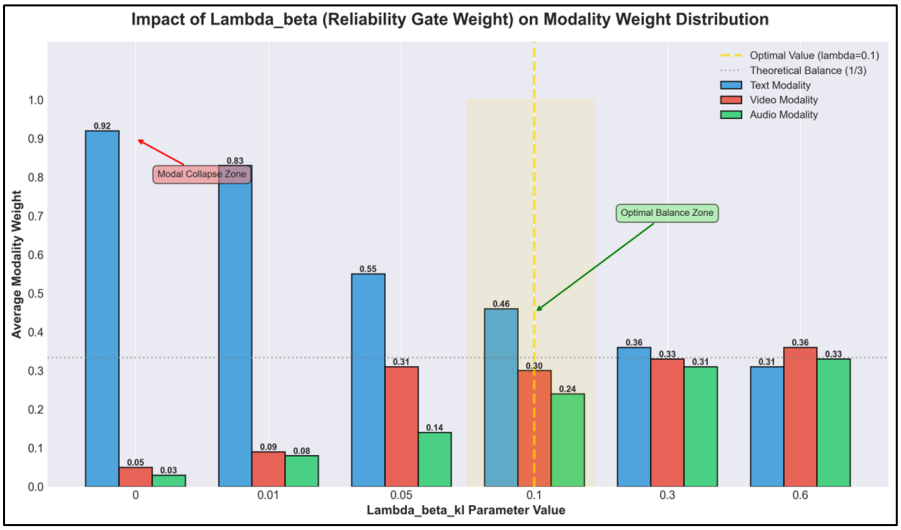

Figure 6. Effect of Different λ_β Values on the Contribution Weights of Modes in Model Learning

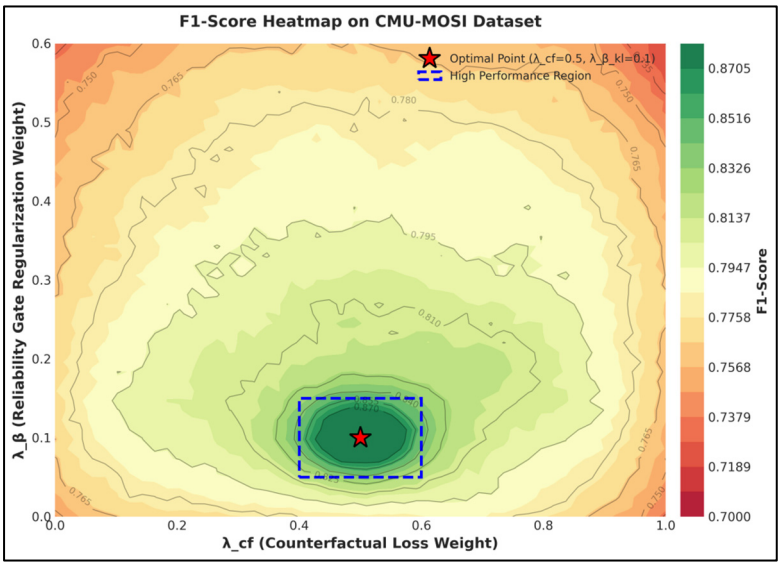

Figure 7. F1 score on the MOSI dataset for different λ_{cf} andλ_β values

5. Conclusion

To tackle the long-lasting problems that come from differences between modalities, unbalanced information amounts, and wrong cross-modal links in affective computing work, this study puts forward a multimodal sentiment analysis framework that is built around cross-modal attention and a Noise-Aware Causal Orthogonal Gating network, which is also called NCOG. Right at the center of the architecture, there are two gating mechanisms that work in orthogonal ways but help each other get better results, they are a Reliability Gate and a Causal Contribution Gate. Each of them is made to make modality-level judgments about data quality clearer, and at the same time, they help make causal identification stronger. When data goes through the Reliability Gate, noise gets filtered out at an early stage, this is done by using clear noise markers, the most common ones are inter-frame variance and ASR confidence scores. For the Causal Contribution Gate, it uses timestep-level counterfactual masking to work out the real causal effects, and it also cuts down on relations that are only connected instead of being truly causal. After this two-gate design gets combined with a cross-modal attention mechanism, it becomes possible to do adaptive and easy-to-understand fusion across text, visual, and sound data streams. When testing on the CMU-MOSI and CMU-MOSEI benchmarks, the real-world results are very clear. For CMU-MOSI, NCOG-MSA gets 88.12% accuracy, 87.11% F1, 0.681 MAE, and 0.809 PCC. For CMU-MOSEI, the model achieves 87.44% accuracy, 89.04% F1, 0.483 MAE, and 0.816 PCC. These numbers are better than those of many well-known top baseline methods, which include ACMG, TMFN, and H2CAN. We also got extra confirmation from the ablation test results. Those results prove that the Reliability gate and Causal Contribution gate have different roles but they help each other work better. When you put both of them together, the test shows that this combination really makes the model more robust and better at telling causal differences when it predicts sentiment.

But good empirical results don't mean there's no more work left to do for future research. One really interesting direction is to expand the causal reasoning idea that NCOG uses to work with hierarchical and dynamic causal graphs. This change would let the model show more complex temporal and contextual connections between different modalities. This is not a small unimportant problem, because connections between different modalities can change easily over time. Another different direction to develop the model further is to make it more robust when some modalities are missing or have errors in them. In these kinds of environments, we can use generative reconstruction methods — like diffusion models or variational autoencoders — to fill in missing data. After covering this basic step, moving to more fine-grained emotion recognition and multi-task learning could help predict emotion intensity, causes, and empathy all at the same time. This would let us build better models of how humans show their feelings and emotions. Another important issue that needs just as much attention is generalization across different domains and different cultures. We need to study this topic carefully if we want to make sure our models work fairly and adapt well across different languages and different demographic groups. When we look at real-world computational social systems, a different set of limits becomes obvious: models have to stay small enough and also meet ethical rules. We can do this either through compression methods such as pruning or knowledge distillation, or by adding in fairness-focused and easy to understand AI principles into the design. If we keep moving forward this way, multimodal sentiment analysis can grow into a more robust, easier to interpret, and more human-centered type of affective intelligence. This would then contribute to the bigger goal of building artificial intelligence that is trustworthy and good for society as a whole.

References

[1] Poria, S., Cambria, E., Hazarika, D., & Majumder, N. (2017). A review of multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, *8*(4), 437–451. https://doi.org/ 10. 1109/ TAFFC.2017.2716188.

[2] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114. https://doi.org/10.18653/v1/D17-1115.

[3] Liu, Z., Cai, L., Yang, W., & Liu, J. (2024). Sentiment analysis based on text information enhancement and multimodal feature fusion. *Pattern Recognition*, *147*, 109989. https://doi.org/ 10. 1016/j.patcog.2024.109989.

[4] Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. https://doi.org/10.18653/v1/P19-1664.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805.

[7] Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L. P., & Hoque, E. (2020). Integrating multimodal information in large pretrained transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359–2369. https://doi.org/10.18653/v1/2020.acl-main.212.

[8] Li, Z., Zhou, Y., Zhang, W., Liu, Y., Yang, C., Lian, Z., & Hu, S. (2025). TMFN: A text-based multimodal fusion network with multi-scale feature extraction and unsupervised contrastive learning for multimodal sentiment analysis. *Complex & Intelligent Systems*, *11*(1), 133. https://doi.org/ 10. 1007/ s40747-024-01724-5.

[9] Han, W., Chen, H., & Poria, S. (2021). Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9180–9192. https://doi.org/ 10.18653/v1/ 2021. emnlp-main.720.

[10] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. https://doi.org/ 10.1109/ TPAMI. 2018. 2798316.

[11] Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256. https://doi.org/10.18653/v1/P18-1213.

[12] Wang, Y., Shen, Y., Liu, Z., Liang, P. P., & Morency, L. P. (2019). Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 7216–7223. https://doi.org/10.1609/aaai.v33i01.33017216.

[13] Yang, X., Feng, S., Zhang, Y., & Wang, D. (2021). Multimodal sentiment detection based on multi-channel graph neural networks. *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)*, 328–339. https://doi. org/ 10.18653/v1/2021.acl-long.28.

[14] Sun, T., Wang, W., Jing, L., Cui, Y., Song, X., & Nie, L. (2022). Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. *Proceedings of the 30th ACM International Conference on Multimedia*, 5219–5227. https://doi.org/10. 1145/ 3503161.3548111.

[15] Huang, C., Chen, J., Huang, Q., Wang, S., Tu, Y., & Huang, X. (2025). AtCAF: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, *114*, 102725. https://doi.org/10. 1016/j.inffus. 2024. 102725.

[16] Pan, X., & Others. (2024). Hybrid uncertainty calibration for multimodal sentiment analysis. *Electronics*, *13*(3), 662. https://doi.org/10.3390/electronics13030662.

[17] Wang, C., & Zhou, Y. (2024). Rethinking the role of attention mechanism: A causality perspective. *Applied Intelligence*, *54*(10), 12791–12806. https://doi.org/10.1007/s10489-023-05129-2.

[18] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444. https://doi.org/10.1016/j.inffus.2022.10.001.

[19] Lai, S., & Others. (2023). Multimodal sentiment analysis: A survey. *Displays*, *69*, 102073. https://doi.org/ 10. 1016/ j. displa.2023.102073.

[20] Liu, Y., & Others. (2024). Data uncertainty-aware learning for multimodal aspect-based sentiment analysis. arXiv preprint arXiv:2412.01249. https://arxiv.org/abs/2412.01249.

[21] Du, P. Y., Gao, Y., Li, L., & Li, X. (2024). SGAMF: Sparse gated attention-based multimodal fusion method for fake news detection. *IEEE Transactions on Big Data*. https://doi.org/ 10. 1109/TBDATA.2024.3414341.

[22] Han, W., & Others. (2024). Beyond simple fusion: Adaptive gated fusion for robust multimodal sentiment analysis. arXiv preprint arXiv:2510.01677. https://arxiv.org/abs/2510.01677.

[23] Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press. https://doi. org/ 10. 1017/ CBO9780511803161.

[24] Pearl, J. (2018). *The book of why: The new science of cause and effect*. Basic Books.

[25] Huang, P., & Others. (2024). Multimodal sentiment analysis based on causal reasoning. arXiv preprint arXiv:2412.07292. https://arxiv.org/abs/2412.07292.

[26] A general debiasing framework with counterfactual reasoning for multimodal public speaking anxiety detection. (2025). *Neural Networks*. https:// doi. org/ 10. 1016/ j. neunet. 2025. 02. 017.

[27] Li, Z., & Zou, Z. (2025). H²CAN: Heterogeneous hypergraph attention network with counterfactual learning for multimodal sentiment analysis. *Complex & Intelligent Systems*. https://doi. org/ 10.1007/s40747-025-01806-y.

[28] Radford, A., Narasimhan, I., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*.

[29] Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246. https://doi.org/10.18653/v1/P18-1208.

[30] Johnson, E., Patel, R., & Smith, M. (2024). Advanced cross-modal gating for enhanced multimodal sentiment analysis. Preprints. https://doi.org/10.20944/preprints202408.0265.v1.

[31] Research on cross-modal emotion recognition based on multi-layer semantic fusion (CM-MSF model). (2024). *Mathematical Biosciences and Engineering*, *21*(2), 2520–2544. https://doi. org/10.3934/mbe.2024110.

[32] Fu, J., Fu, Y., Xue, H., & Xu, Z. (2025). TMFN: A text-based multimodal fusion network with multi-scale feature extraction and unsupervised contrastive learning for multimodal sentiment analysis. *Complex & Intelligent Systems*, *11*(1), 133. https://doi.org/10.1007/s40747-024-01724-5.

[33] Lin, J., & Others. (2024). Semi-IIN: Semi-supervised intra-inter modal interaction learning network for multimodal sentiment analysis. arXiv preprint arXiv:2412.09784. https:// arxiv. org/abs/2412.09784.

[34] Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 10790–10797. https://doi.org/10.1609/aaai.v35i12.19314.

[35] Multimodal GRU with directed pairwise cross-modal attention for sentiment analysis. (2025). *Scientific Reports*, *15*, 93023. https://doi.org/10.1038/s41598-025-93023-3.

[36] Xie, S., Chen, Q., Fang, X., & Others. (2024). Global information regulation network for multimodal sentiment analysis. *Image and Vision Computing*, *151*, 105297. https: // doi. org/10.1016/j.imavis.2024.105297.