Journal of Artificial Intelligence Practice (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2026.090112
ISSN 2371-8412 Vol. 9 Num. 1

Emotion Causal Chain Analysis Method Based on Multi-Modal Feature Fusion

Lin Gan, Zhengpeng Zhang

School of Information and Intelligent Engineering, University of Sanya, Sanya, Hainan, 572100, China

Keywords: Multi-modal Fusion; Emotion Causal Chain; Counterfactual Learning; Causal Inference; Sentiment Analysis

Abstract: The development of multi-modal large models provides a robust representation foundation for sentiment analysis. However, current research primarily focuses on static classification tasks, neglecting the dynamic evolutionary nature of emotions. This paper centers around the core concept of the "emotion causal chain," systematically reviewing the research status of multi-modal feature fusion and counterfactual learning in the field of sentiment analysis. Based on defining the theoretical connotation of the emotion causal chain, it critically compares early fusion, attention mechanisms, graph neural networks, and large model fusion paradigms from a technological evolution perspective, pointing out the fundamental limitations of existing methods in terms of interaction range, spurious correlation control, and interpretability. Furthermore, it focuses on discussing the theoretical foundation and application paths of counterfactual learning, elucidating its methodological advantages in modal decoupling, temporal intervention, and path identification. It also systematically summarizes the implementation framework based on generative models, contrastive learning, causal attention, and large model integration. The research suggests that counterfactual learning enables models to go beyond statistical associations and touch upon the causal mechanisms of emotion evolution, enhancing interpretability and robustness while providing computational tools for analyzing cross-modal emotion transmission paths. Finally, it looks forward to future directions, emphasizing that building a multi-modal fusion model with causal reasoning ability is the key to achieving interpretable and strongly generalizable emotional intelligence.

1. Introduction

Human emotion is essentially the product of multimodal collaborative expression. Linguistic content conveys cognitive evaluations, vocal prosody carries emotional intensity, facial expressions and body movements reveal inner states, and physiological signals reflect arousal levels[1]. A single modality can only present a partial dimension of emotion. Only cross-modal integration can approximate the full picture of emotional perception. In recent years, the development of multimodal large models has brought about a paradigm shift in emotion analysis. Architectures represented by CLIP, Flamingo, and GPT-4V have overcome the technical bottleneck of aligning heterogeneous modal representations. Transformers and their variants enable cross-modal

interaction to be realized in deep networks, providing a powerful computational framework for capturing modal complementarity and redundancy in emotional expressions. This evolution has propelled emotion analysis from shallow feature engineering to deep representation learning, and from unimodal independent modeling to multimodal collaborative understanding[2].

Although multimodal large models have significantly improved the accuracy of emotion recognition, current research still faces fundamental limitations: most models focus on the mapping from input to emotion labels, simplifying emotions into discrete categories or continuous-dimensional classification tasks, while neglecting the dynamic evolutionary nature of emotions. Emotions in real-world scenarios are not static. Instead, they are triggered by specific causes, unfold on a timeline, are transmitted between modalities, and may be transformed into causal chains of subsequent emotional states[3]. For example, in a dialogue, a slight offense in the content of the speech (cause) may trigger micro-expression changes in the listener (effect), which in turn becomes the trigger for subsequent verbal retaliation. Understanding this "emotional causal chain"-the complete dynamic process from emotional elicitation, representation, transmission to transfer-is crucial for building truly intelligent emotional interaction systems. Therefore, the focus of research needs to shift from "what emotion is it" to "why is the emotion generated" and "how does the emotion evolve," that is, from emotion recognition to emotion causal understanding.

To lay the foundation for the full text discussion, three core concepts are defined:

Multimodal feature fusion refers to the process of aligning and integrating data from heterogeneous modalities such as text, audio, visual, and physiological signals to form a joint representation. The fusion level can be divided into early fusion (feature-level concatenation), late fusion (decision-level voting), and hybrid fusion (multi-level interaction). The core challenge lies in solving the heterogeneity, asynchrony, and complementarity between modalities[4].

The emotional causal chain is defined as the complete dynamic causal structure of emotion from elicitation, representation, transmission to transfer. Unlike static emotion classification, the emotional causal chain focuses on the dependencies between emotional elements: what external stimuli or internal cognition (cause) leads to a specific emotional reaction (effect); and how this emotion becomes the cause of subsequent emotional states or behavioral expressions. This concept draws on the ideas of structural causal models to advance emotion understanding from correlation analysis to causal inference.

Counterfactual learning originates from counterfactual reasoning in causal inference. The core is to answer the question of "how will the result change if a certain condition changes." In emotion analysis, counterfactual learning enables the model to learn to distinguish between true emotional causal clues and superficial false associations by constructing counterfactual samples corresponding to factual samples (such as replacing emotional words, modifying facial expressions, and changing vocal prosody), thereby achieving modality decoupling, path identification, and explainability enhancement.

2. Technical Genealogy and Causal Modeling Capability Analysis of Multimodal Emotion Fusion Methods

2.1 Classification Framework and Evolutionary Logic of Fusion Paradigms

Multimodal feature fusion is the cornerstone of emotional causal chain analysis, and its technological evolution reflects the deepening understanding of emotional expression complexity in affective computing. Based on the level and interaction mechanism of fusion, existing methods can be summarized into four progressive stages.

2.1.1 Early Fusion, Late Fusion, and Hybrid Fusion

Early fusion integrates raw inputs at the feature level, combining heterogeneous modal feature vectors into a joint representation through concatenation, addition, or outer product, and inputs them into a single model for end-to-end learning. This paradigm assumes that the interaction between modalities can be fully captured at the input stage. Its advantage lies in utilizing the complementarity of underlying features, but it faces problems such as difficult modal asynchronous alignment, dimensionality disasters caused by feature heterogeneity, and insufficient robustness in modal missing scenarios[5]. Late fusion independently processes each modality, trains single-modal models separately, and integrates classification results at the decision level through weighted voting, averaging, or meta-learning. This method has high computational efficiency and natural robustness to modal missing, but it sacrifices the possibility of cross-modal interaction modeling, making it difficult to capture the subtle modal synergy effects in emotional expression. Hybrid fusion attempts to strike a balance between the two by designing multi-level interaction paths and introducing cross-modal information in the feature extraction and decision-making stages, laying the foundation for more complex interaction mechanisms in the future.

2.1.2 Interactive Fusion Based on Tensors and Attention

In order to break through the representational limitations of simple fusion on modal interaction, the Tensor Fusion Network (TFN) introduces three-dimensional outer product operations to explicitly model pairwise and three-way interactions between the three modalities. This method generates high-dimensional tensors by calculating the Cartesian product of modal feature vectors to capture multiplicative interactions between modalities, but its computational complexity increases exponentially with the number of modalities, and the outer product operation lacks interpretability. The introduction of the attention mechanism marks a qualitative change in the fusion paradigm. Cross-modal Attention enables the model to dynamically align feature sequences of different modalities at the time step level. Multimodal Transformer (MuIT) implements temporal interaction of non-aligned modal sequences through directional pairwise cross-modal attention modules. The self-attention mechanism enables joint modeling of dependencies within and between modalities in the same framework, providing technical support for capturing long-range dependencies in emotional dynamic evolution.

2.1.3 Structured Fusion Based on Graph Neural Networks and Hypergraphs

Emotional expressions in dialogue scenarios naturally have graph-structured characteristics-each modal feature constitutes a node, and temporal adjacency and semantic similarity constitute edges. Graph neural networks explicitly model temporal dependencies within modalities and interaction relationships between modalities through a message-passing mechanism. However, standard graph structures are limited to binary edges, which can only represent pairwise interactions, making it difficult to handle higher-order relationships involving more than three emotional elements in multimodal emotional expressions. The introduction of hypergraph neural networks breaks through this limitation: hyperedges can connect any number of nodes, enabling the model to simultaneously capture the multiple synergies of text content, speech prosody, facial expressions, and contextual information[6]. The cross-modal hypergraph neural network proposed by Jiang Kun et al. uses a hypergraph structure to connect multiple modal nodes, fully mining high-order emotional interaction information within and between modalities, and adaptively optimizes the hypergraph structure through a point-edge cross-attention mechanism, pruning redundant connections to enhance the accuracy of emotional representation.

2.2 Theoretical Requirements

2.2.1 Temporal Dependency

Emotions are not instantaneous, discrete events but rather dynamic processes that unfold on a timeline. The elicitation of a specific emotion often stems from the accumulation of prior events, and the current emotional state becomes the cause of subsequent emotional reactions. This temporal causal structure requires fusion methods to possess the ability to model long-range temporal dependencies-not only to identify the emotional label at the current moment but also to trace the trigger points and transmission paths of emotional changes. While positional encoding in Graph Neural Networks and Transformers partially addresses the issue of temporal modeling, most methods are still limited to capturing temporal patterns of correlation rather than temporal dependencies in a causal sense. The Hierarchical Emotion Causal Graph Model (HE-CGM) proposed by Ang et al., which captures global dependencies through a structured causal matrix and models local state transitions through a recurrent module, represents a significant exploration toward temporal causal modeling.

2.2.2 Inter-Modality

The expression of emotions across different modalities exhibits temporal misalignment and semantic complementarity-facial micro-expressions may reveal true emotions before speech, and vocal prosody may modify the surface semantics of text content. Emotional causal chains require fusion methods to be able to parse this cross-modal transmission mechanism: How does the emotional signal of one modality trigger reactions in other modalities? Does a causal direction exist in the inter-modal modulation? Although attention mechanisms can calculate cross-modal association weights, these weights only represent statistical correlations and cannot distinguish causal directions. Disentangling the true causal transmission paths between modalities requires fusion methods to have intervention capabilities-by changing the input of one modality and observing the changes in the representation of another modality, rather than merely calculating their similarity[7].

2.2.3 Confounding Factors

Emotional expressions in real-world scenarios are subject to interference from multiple confounding factors: the context of the conversation, cultural background, individual expression habits, etc., which can simultaneously affect multi-modal signals and emotional labels, leading to spurious correlations. For example, in certain cultures, a smile may express embarrassment rather than pleasure, and an individual's habitually exaggerated gestures may be unrelated to the intensity of emotion. Traditional fusion models lack explicit control over confounding factors, making them highly susceptible to learning contextual biases rather than genuine emotional causal mechanisms. Emotional causal chain modeling requires fusion methods to have the ability to disentangle-to separate emotion-related causal representations from confounding factors such as context and individual differences, enabling the model to answer counterfactual questions.

2.3 Limitations of Existing Fusion Methods in Causal Chain Modeling

2.3.1 Limitations in Interaction Scope

Mainstream cross-modal attention mechanisms are essentially pairwise interactions between modalities-pairwise computation of attention weights between text-audio, text-visual, and

audio-visual, followed by information integration through weighted summation. This stacking of pairwise interactions makes it difficult to capture high-order emotional semantics that emerge only through the collaboration of three or more modalities. Take the sarcastic expression as an example, its recognition often relies simultaneously on the literal meaning of the text, the ironic prosody of the voice, and the contradictory expressions on the face. Although pairwise interaction models can calculate the correlation between text and voice, and text and expression, respectively, they cannot model the nonlinear coupling relationship between the three. The introduction of hypergraph networks partially alleviates this problem, but how to define the semantic rationality of hyperedges and how to learn dynamically changing hypergraph structures remain open challenges.

2.3.2 Inherent Problems of Spurious Correlations

Multimodal models generally exhibit modality bias-a tendency to rely on one easily learned modality while ignoring emotional cues in other modalities. The root of this bias lies in the confusion between statistical correlation and causality: after the model discovers that textual sentiment words are highly correlated with labels, it no longer invests computational resources in learning the more subtle emotional signals in the audio and visual modalities. A more fundamental problem is that the dataset itself encodes various biases-consistency of the acquisition context, the cultural background of the annotators, and the uneven distribution of emotional categories-all of which may cause the model to learn spurious modality-label associations. When the model establishes a strong correlation between "smile" and "happiness," it cannot distinguish whether this correlation stems from a genuine emotional expression or is due to the fact that most smiling samples in the dataset were indeed collected in happy situations[8].

2.3.3 Lack of Explainability

The core requirement of explainable emotion computation is to understand "why" the model makes a specific emotional judgment. However, current mainstream fusion methods are essentially black-box models-their internal representations are high-dimensional, nonlinear, and distributed encodings, making it difficult to map them back to understandable causal factors. Attention weights are often used as explanatory tools, but attention only reflects statistical patterns of information allocation, not evidence of causal relationships. To answer the core question of the causal chain-"which feature of which modality triggered the emotional shift at which point in time?"-existing fusion methods lack corresponding representation decoupling mechanisms and counterfactual traceability. This lack of explainability not only hinders the academic understanding of emotional mechanisms but also limits the practical application of emotional computation in high-risk areas.

3. Counterfactual Learning: A Methodological Breakthrough in Modeling Emotional Causal Chains

3.1 Theoretical Foundation

Counterfactual reasoning stems from the core idea of causal inference, aiming to answer the question of "what would happen to the result if a certain condition changed?" This ability is regarded as an important component of general artificial intelligence because it enables machines to go beyond the superficial associations of observed data and touch upon the causes and consequences of how things develop and change.

3.1.1 The Three Rungs of Causal Inference

Judea Pearl divides causal inference into three progressive rungs. The first rung, "Association," answers "what do I see?"-that is, discovering patterns through statistical correlations. The second rung, "Intervention," answers "what if I do?"-observing changes in results by actively changing variables. The third rung, "Counterfactual," answers "what if?"-retrospectively imagining possible outcomes in alternative scenarios. Traditional multimodal sentiment analysis has long remained at the first rung-the model learns the statistical correlations between modal features and sentiment labels but cannot answer counterfactual questions such as "if a certain segment of speech is deleted, will the sentiment judgment change?" Parsing emotional causal chains requires the model to leap to the third rung, possessing counterfactual reasoning ability.

3.1.2 Structural Causal Models and Potential Outcome Frameworks

The realization of counterfactual reasoning relies on two major theoretical frameworks. The Structural Causal Model (SCM) represents a causal system as a directed acyclic graph, where each variable is determined jointly by its parent nodes, noise variables, and a deterministic function; intervention operations correspond to the replacement of causal mechanisms. The Potential Outcome Framework, on the other hand, defines the potential outcome of each individual under a specific treatment, and the comparison between factual and counterfactual outcomes constitutes the estimation of the causal effect. The two frameworks are equivalent; the former expresses the causal graph structure more intuitively, and the latter quantifies the causal effect more precisely. In multimodal sentiment analysis, SCM can be used to model the causal transmission paths between modalities, and the potential outcome framework can be used to estimate the causal contribution of a specific modality to sentiment judgment.

3.1.3 The Generation Logic of Counterfactual Samples

The generation of counterfactual samples follows the minimal intervention principle-changing only the target variable to observe the resulting change while keeping other conditions unchanged. This principle ensures that the difference between counterfactual and factual samples originates only from the intervened variable, thereby isolating the causal effect. In sentiment analysis, counterfactual samples can be generated by replacing sentiment words, modifying facial expressions, or changing speech prosody, enabling the model to learn to distinguish between genuine emotional causal clues and superficial false associations.

3.2 Application Paths of Counterfactual Learning in Multimodal Sentiment Analysis

3.2.1 Counterfactual Disentanglement at the Modality Level

Multimodal models commonly exhibit modality bias-a tendency to rely on easily learned modalities while neglecting emotional cues from other modalities. Counterfactual disentanglement addresses this by constructing counterfactual interventions at the modality level, compelling the model to focus on the true causal contributions of each modality. The H₹CAN model designs a Counterfactual Intervention Task (CIT), generating attention scores through a Gaussian distribution and attaching weights to biased modalities, amplifying the impact of biased information. This maximizes the predictive differences between factual and counterfactual branches, guiding attention learning to focus on causal relationships between modalities. This method uses single-modality labels to adaptively identify biased modalities, effectively mitigating the varying degrees of influence from different modality biases.

3.2.2 Counterfactual Intervention in the Temporal Dimension

The evolution of emotions over time possesses a causal structure-the current emotional state is the result of previous events and the cause of subsequent reactions. Counterfactual intervention in the temporal dimension involves altering the modal input at specific time points and observing changes in subsequent emotional representations, thereby identifying key nodes in emotional transmission. This approach requires the model to have temporal causal modeling capabilities, distinguishing between temporal correlation (succession) and temporal causality (the former causing the latter). Comparing the counterfactual sequence after intervention with the original sequence can reveal the core drivers of emotional evolution.

3.2.3 Counterfactual Identification of Emotion Transmission Paths

Cross-modal emotion transmission is the core mechanism of the emotion causal chain-changes in speech prosody may stem from the emotional load of textual content, and changes in facial expressions may be triggered by speech intonation. Counterfactual identification infers the causal direction between modalities by blocking or enhancing the input of specific modalities and observing the responses of other modal representations. For example, if keeping the text constant while replacing the speech with a neutral tone leads to a change in emotional judgment, it proves that the speech modality has an independent causal contribution. If changing the text but keeping the speech constant does not change the judgment, it indicates that the emotion of the sample mainly originates from speech cues. This counterfactual intervention shifts the relationship between modalities from statistical correlation to causal attribution.

3.3 Implementation Methods and Technical Framework for Counterfactual Learning

3.3.1 Counterfactual Sample Construction Based on Generative Models

The quality of counterfactual samples directly determines the effectiveness of causal learning. Generative adversarial networks, variational autoencoders, and diffusion models are widely used to generate realistic and reliable counterfactual samples. The DiCap model utilizes a diffusion process to iteratively sample gradients from the marginal and conditional distributions of the causal model, generating counterfactual samples that satisfy the minimum sufficiency criterion. It also theoretically guarantees the identifiability of counterfactual results and the boundedness of estimation errors. In sentiment analysis, diffusion models can be used to generate realistic multimodal data that retains identity features and contextual information while changing emotional expressions, providing high-quality training material for causal learning.

3.3.2 Counterfactual Representation Learning Based on Contrastive Learning

Contrastive learning learns discriminative representations by pulling similar samples closer and pushing dissimilar samples further apart. Introducing counterfactual samples into the contrastive learning framework allows the model to learn to distinguish between causally related features and superficially associated features. Specifically, factual samples and their counterfactual counterparts constitute hard negative examples-they are highly similar in shallow features (e.g., same identity, same context) but have different sentiment labels due to changes in interventional variables. The model is forced to learn to capture true causal clues rather than relying on confounding factors such as identity and context. The DiCap model integrates generated counterfactual samples into a contrastive learning framework, achieving the extraction of prompt representations that are precisely aligned with the causal features of the data.

3.3.3 Counterfactual Intervention Based on Causal Attention Mechanism

Attention mechanisms are widely used to explain model decisions, but attention weights only reflect statistical associations, not causal evidence. Causal attention mechanisms place attention learning under the constraints of causal relationships by introducing counterfactual interventions. Counterfactually Decoupled Attention Learning (CDAL) explicitly models the causal relationship between attention visual traces and source model attribution. It separates discriminative model-specific artifacts and confounding source biases through counterfactual decoupling, maximizing causal effects to guide the network to capture essential patterns that generalize to unseen samples. The counterfactual intervention task in H ℃AN also utilizes causal inference to maximize the prediction difference between the factual branch and the counterfactual branch, guiding hypergraph attention learning to focus on real emotional interactions.

3.3.4 Combination Strategies of Counterfactual Reasoning and Multimodal Large Models

Multimodal large models possess powerful representation and cross-modal alignment capabilities, but their black-box nature exacerbates the difficulty of causal mechanism analysis. Combining counterfactual reasoning with large models can break through from two directions: first, introducing counterfactual objectives in the pre-training stage to enable the model to distinguish between causally related and spuriously correlated features; second, explaining model decisions through counterfactual intervention in the inference stage. The former requires constructing large-scale counterfactual pre-training data or designing counterfactual contrastive losses, while the latter requires embedding interventional causal modules in the model architecture. Current research is still in its early stages, but the deep integration of counterfactual reasoning and large models is expected to achieve a new generation of sentiment analysis models that combine representation power with causal explainability.

4. Conclusion

This paper systematically reviews the evolution of multimodal fusion techniques and the methodological breakthroughs in counterfactual learning, centering on the core issue of "emotion causal chain analysis based on multimodal feature fusion." The research reveals that the paradigm of emotion analysis is undergoing a profound shift from static classification to dynamic causal understanding. This shift places triple theoretical demands on fusion methods: temporal dependency, inter-modal nature, and control of confounding factors. Although existing fusion methods have continuously evolved in interaction depth, the limitations of pairwise interaction scope, the inherent problem of spurious correlations, and the fundamental lack of interpretability make it difficult to access the causal mechanisms of emotional evolution. The introduction of counterfactual learning provides methodological support for overcoming the above limitations-through modal decoupling, temporal intervention, and path identification, the counterfactual framework enables the model to distinguish between causal clues and surface associations, enhancing interpretability and robustness while revealing the internal paths of cross-modal emotion transmission. However, the authenticity of counterfactual samples, computational complexity, and the lack of evaluation criteria remain core challenges restricting its development. Future research needs to continue to break through in three directions: temporal causal modeling, external knowledge fusion, and evaluation system construction, to ultimately achieve emotion intelligence with both representation ability and causal interpretability.

Acknowledgement

This work was supported by the Education Department of Hainan Province. Grant No. Hnky2026-35 for the project entitled Research on Mental Health Analysis Based on Multimodal Large Models.

References

[1] Jie Liliang, Zou Yangmeng, Li Zhengxiu, et al. Method for Emotion Conversion Recognition Based on Cross-modal Feature Fusion and Global Perception [J]. Journal of Biomedical Engineering, 2025, 42(05): 977-986.
[2] Zhang Zhiwen, Yu Nai Gong, Bian Yan, et al. Research on Emotion Recognition Based on Multi-modal Physiological Signal Feature Fusion [J]. Journal of Biomedical Engineering, 2025, 42(01): 17-23.
[3] Zhang Yiqing. Research on Emotional Understanding Method Based on Multi-modal Feature Fusion [D]. Beijing University of Posts and Telecommunications, 2025.
[4] Gao Jingjing. Research on Multi-modal Emotion Recognition Based on Multi-dimensional Features and CRNN [D]. Qufu Normal University, 2025.
[5] Wang Ke. Research on Emotion Recognition Method Based on Multi-modal Fusion [D]. Jilin University, 2025.
[6] Zheng Peng. Research on Audio-Video Emotional Analysis Based on Multi-modal Feature Fusion [D]. Jiangxi Normal University, 2025.
[7] Fu Junsong. Research on Emotional Analysis Based on Multi-modal Feature Fusion [D]. Chongqing University of Technology, 2025.
[8] Yang Liang. Research on Emotional Analysis Based on Multi-modal Feature Fusion [D]. Guangxi Normal University, 2025.