Beyond Correlation-Driven Fusion: Sample-Adaptive Causal Evidence Learning for Multimodal Sentiment Analysis

Anonymous EMNLP submission

Abstract

Existing multimodal sentiment analysis (MSA) methods rely on representation learning and cross-modal fusion. However, many of them follow a correlation-driven paradigm that may encode lexical shortcuts, visual biases, and non-causal correlations into predictions. To address this problem, we propose a Consistency-Conflict Fusion with Sample-Adaptive Causal Evidence (CCF-SACE) framework, revisiting MSA from the perspective of representation-level causal modeling. First, we use off-the-shelf encoders to extract token-level textual features and patch-level visual features, and employ an adaptor to project them into a shared semantic space. The framework then performs sample-adaptive cross-modal causal evidence selection, where textual causal evidence is selected under visual context and visual causal evidence under textual context. The remaining irrelevant contexts are modeled as non-causal proxies, allowing the model to separate emotion-supportive evidence from shortcut-related information within each sample. After obtaining such sample-specific causal evidence, CCF-SACE introduces a covariance-based consistency-conflict-aware fusion module to avoid directly fusing all selected cues, constructing an evidence-level mediator by modeling cross-modal consistency and discrepancy. Moreover, since the residual proxy stream may still carry label-correlated shortcut information, we introduce non-causal proxy suppression and counterfactual sensitivity objectives to weaken its standalone predictability and enlarge the discrepancy between evidence-mediated and proxy-only predictions. Experiments on five benchmarks show consistent improvements over state-of-the-art methods, especially on fine-grained affective recognition.

1 Introduction

Multimodal sentiment analysis aims to infer human sentiment tendencies from heterogeneous information such as text and images. This task plays an important role in affective computing and multimodal understanding (Das and Singh, 2023; Gandhi et al., 2023; Singh et al., 2024). Compared with unimodal sentiment analysis, multimodal sentiment analysis can exploit explicit semantic expressions in text and visual sentiment cues in images, thereby providing richer contextual information for sentiment judgment (Niu et al., 2016; Xu and Mao, 2017; Liu et al., 2024).

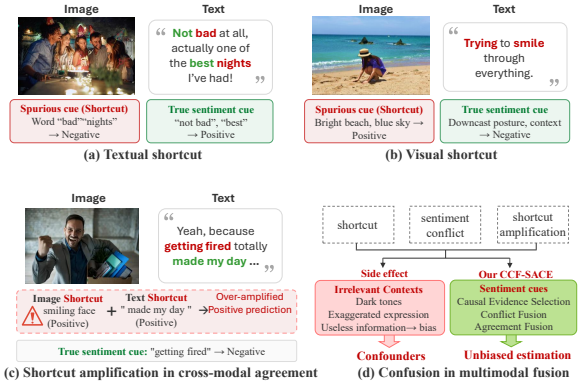

Figure 1: (a) Sentiment words mislead the prediction; (b) appearance bias misleads the prediction; (c) both modalities share shortcuts, leading to over-confident errors; (d) the model over-relies on one modality and ignores the other.

Recent studies have made substantial progress in multimodal representation learning and cross-modal fusion (Zhu et al., 2022; Li et al., 2022; Yang et al., 2023; Wu et al., 2024). Typical approaches include attention mechanisms, disentangled representation learning, and semantic enhancement methods based on prompts or reasoning chains (Ling et al., 2022; Yu et al., 2022; Wang et al., 2024; Pang et al., 2024; Huang et al., 2024). Their common goal is to strengthen the interaction between text and images so that models can capture richer cross-modal correlations. Although these methods achieve strong performance on multiple benchmark datasets, they often rely on an implicit assumption: more sufficient multimodal interaction

leads to more reliable sentiment representations. This assumption does not always hold under real-world or biased data distributions (Liu et al., 2024; Wu et al., 2024). On the textual side, models may rely on high-frequency sentiment words, lexical co-occurrence patterns, or template-like expressions (Sun et al., 2023; Ren et al., 2026). As shown in Fig. 1(a), words such as "bad" and "night" may highly co-occur with negative sentiment classes in the training set, but their true polarity can be affected by context, negation, or discourse shifts. Similar shortcuts also exist in images. Fig. 1(b) illustrates that the model may use bright scenes and specific backgrounds as shortcut cues for predicting positive sentiment. However, these visual factors do not necessarily constitute stable evidence for sentiment judgment.

In other words, both text and images contain genuinely useful evidence for sentiment reasoning as well as non-causal spurious information induced by the data distribution. More importantly, complex cross-modal interactions do not necessarily remove spurious correlations. instead, they may reinforce them (Ma, 2025; Jiang et al., 2025; Chen et al., 2025a). As show in Fig. 1(c), when a text with positive words like "made my day" is paired with a bright smiling face image, the fusion model shows higher confidence in incorrect predictions due to the co-occurrence of textual and visual shortcut cues. However, when the shortcuts in the text and the image conflict with each other or even perform false enhancement, simple fusion may cause the model to overly rely on the shortcuts in the dominant modality while ignoring the crucial evidence provided by the other modality, as shown in Fig. 1(d). Therefore, the core challenge of image-text sentiment analysis is not merely how to enhance multimodal interaction, but how to prevent cross-modal fusion from amplifying non-causal shortcuts, moving from fusing all correlated information toward fusing reliable sentiment evidence. Thus, our goal is to identify causal evidence and perform consistency conflict-aware fusion, so that the model relies on reliable emotional cues rather than misleading shortcuts.

Based on the above analysis, we proposes a Consistency-Conflict Fusion with Sample-Adaptive Causal Evidence (CCF-SACE) to improve the robustness of multimodal sentiment analysis from the perspective of causal evidence modeling. The core idea is that stable sentiment prediction should not directly depend on raw multimodal inputs affected by confounding, but should rely on sentiment-supportive evidence extracted from text and images. To this end, we first design a sample-adaptive cross-modal causal evidence selection mechanism, which evaluates the relevance of textual tokens conditioned on visual context and the relevance of visual patches conditioned on textual context. In this way, each modality is decomposed into causal sentiment evidence and non-causal proxy information. Considering that text and images may exhibit consistent, complementary, or conflicting relations, we further design a consistency-conflict-aware fusion module to construct a representation-level causal evidence mediator from textual and visual causal evidence. The model adaptively fuses causal information according to the degree of consistency among cross-modal sentiment evidence, leading to robust multimodal sentiment analysis. The main contributions are summarized as follows:

(1) We revisit multimodal sentiment analysis from a representation-level causal evidence learning perspective, formulating robust sentiment prediction as an evidence-mediated process. Instead of directly fusing all correlated multimodal information, the proposed framework relies on sentiment-supportive causal evidence for prediction.

(2) A sample-adaptive cross-modal causal evidence selection mechanism is designed to separate causal sentiment evidence from non-causal proxy information. It selects textual evidence under visual context and visual evidence under textual context, thereby dynamically filtering shortcut-related information within each sample.

(3) We construct a consistency-conflict-aware causal fusion framework that adaptively balances agreement-based fusion and conflict-aware evidence weighting, while suppressing non-causal proxy information to reduce reliance on spurious correlations. Extensive experiments on four multimodal sentiment analysis datasets and one multimodal emotion recognition dataset demonstrate the effectiveness and robustness of CCF-SACE.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis has been widely studied from the perspectives of representation learning and cross-modal fusion (Wang et al., 2014; Xu and Mao, 2017; Cheema et al., 2021; Das and Singh, 2023; Wu et al., 2024). Early image-text

methods combine multimodal features with semantic networks, attention, or hierarchical fusion strategies (Xu, 2017; Peng et al., 2021; Hu and Yamamura, 2022). Later studies introduce graph-based interaction, vision-language pre-training, contrastive learning, and cross-modal translation to improve representation quality and modality alignment (Yang et al., 2021a; Ling et al., 2022; Li et al., 2022; Yang et al., 2023; Zhang et al., 2024). These methods have advanced image-text sentiment analysis by enabling more effective information exchange between modalities.

Recent studies further explore modality relationships such as consistency, complementarity, and conflict (Liu et al., 2024; Wu et al., 2024, 2026). They usually adjust modal contributions through shared-private representation learning, adaptive weighting, modality enhancement, feature recovery, or discrepancy-aware fusion (Zhu et al., 2022; Zhang et al., 2024; Wang et al., 2025). These designs show the importance of relationship-aware fusion for multimodal sentiment prediction. However, most existing methods perform relationship modeling on raw features, intermediate features, or global fused representations.

In contrast, our method first selects sentiment-supportive evidence and then performs consistency-conflict modeling on the selected evidence. The key difference is that existing fusion models mainly decide how to fuse modalities, whereas CCF-SACE first decides what evidence should be fused.

2.2 Multimodal Causal Learning

Causal inference has been widely used to mitigate bias and spurious correlations in vision-language and multimodal learning (Yang et al., 2021b; Liu et al., 2024; Meng et al., 2025; Zhu et al., 2025), thereby improving model robustness in tasks such as visual question answering, image captioning, and multimodal sentiment analysis. (Li et al., 2023) proposed a deconfounded causal inference network for image-text matching, which applies backdoor adjustment to mitigate intra-modal and inter-modal spurious correlations. (Chen et al., 2025b) proposed a cross-modal causal relation alignment framework for video question grounding, which aligns question-answer semantics with causal video segments. (Huang et al., 2025) introduced an attention-based causality-aware fusion network for multimodal sentiment analysis by combining front-door text debiasing with counterfactual cross-modal attention. (Chen et al., 2024) proposed a counterfactual multimodal sentiment analysis framework, which estimates and subtracts unimodal direct effects to mitigate modality bias and obtain more unbiased image-text sentiment predictions. (Liu et al., 2025) proposed DPCI for multimodal aspect-based sentiment classification, which combines dual-path counterfactual generation, sample-aware selection, and confidence-guided fusion to mitigate aspect–sentiment spurious correlations. However, existing causal debiasing methods mainly focus on global biases at the dataset level or modality level, lacking modeling of fine-grained causal evidence within individual samples. We introduces a front-door-inspired causal evidence mediation during the representation learning stage. It emphasizes sample-level causal evidence selection and non-causal proxy suppression, making it more suitable for robust sentiment reasoning in image-text sentiment analysis.

3 Methodology

We define the multimodel sentiment analysis task and formulate the structural causal view of the problem. Then, we explain how the proposed method approximates front-door deconfounding at the representation level. After that, we provide a detailed introduction to our proposed network. Fig. 2 illustrates the overall architecture of CCF-SACE.

3.1 Task Definition

Given a training set $\mathcal{D} = \{(X_t^i, X_v^i, y_i)\}_{i=1}^N$, where X_t^i and X_v^i denote the textual and visual inputs and $y_i \in \{1, \ldots, C\}$ denotes the sentiment label, multimodal sentiment analysis aims to learn a robust predictor from an image-text pair to its sentiment category. In general, a conventional model directly encodes and fuses the multimodal sample and then maps the fused representation to the corresponding sentiment category, which can be written as:

$$p(y \mid X_t, X_v) = p(y \mid F), F = \mathrm{Fuse}(X_t, X_v). \tag{1}$$

Such a direct fusion path may absorb lexical shortcuts, visual biases, and other spurious cross-modal correlations when the training data contain modality-specific confounding factors.

3.2 Representation-Level Causal Deconfounding

Let C_t and C_v denote latent textual and visual confounders. They may simultaneously influence observed multimodal inputs and sentiment labels, in-

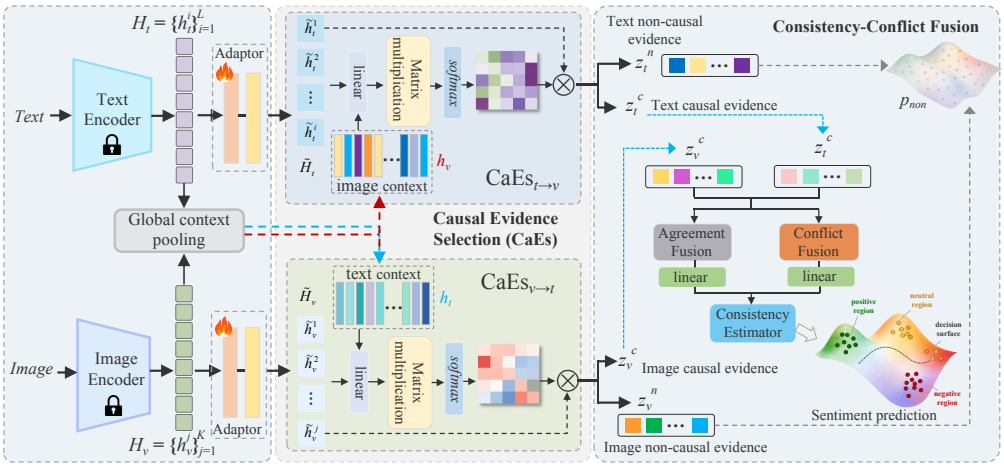

Figure 2: Overall architecture of CCF-SACE. Given an image–text pair, encoders extract textual token and visual patch features, which are projected by a two-layer adaptor into a shared semantic space. Causal Evidence Selection then selects textual evidence under visual context and visual evidence under textual context. Finally, Consistency–Conflict Fusion constructs a covariance-aware evidence mediator to balance agreement and conflict branches for robust sentiment prediction.

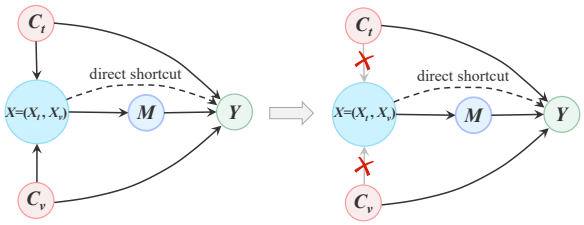

Figure 3: Front-door-inspired causal structure for multimodal sentiment analysis.

ducing shortcut paths that make correlation-driven fusion unreliable. Fig. 3 shows the causal relationships among the textual feature X_t, the visual feature X_v, textual confounder C_t and visual confounder C_v. The confounded paths $X_t \leftarrow C_t \rightarrow Y$ and $X_v \leftarrow C_v \rightarrow Y$ affect the prediction of Y. To alleviate this problem, we introduce an evidence-level mediator M between the multimodal inputs and predictions. The M is constructed from textual causal evidence Z_t^c and visual causal evidence Z_v^c, while the remaining information is treated as non-causal proxy information Z_t^n and Z_v^n.

Inspired by the front-door view (Pearl et al., 2016), we build a causal mediator at the representation layer. Assuming $X = (X_t, X_v)$ and M mediates the stable effect from X to Y, the causal effect can be written as:

$$p(Y \mid do(X = x)) = \sum_m p(m \mid x) \sum_{x'} p(Y \mid m, x')p(x'). \tag{2}$$

Ideally, the causal effect of multimodal inputs on the sentiment label can be estimated through the mediator M, while the influence of confounders in the original input space is weakened by marginalizing over the input distribution. In practice, however, the true causal mediator is not directly observable, and all possible text-image input combinations cannot be enumerated. Therefore, this paper performs a representation-level approximation of front-door adjustment. Specifically, $p(m \mid x)$ in the classical front-door formula is jointly approximated by sample-adaptive causal evidence selection and consistency-conflict fusion, while the inner marginalization term $\sum_{x'} p(Y \mid m, x')p(x')$ cannot be computed exactly in the high-dimensional image-text space. We therefore approximate it at the representation level using a prediction that depends only on the causal evidence mediator. In this formulation, M is mainly constructed from causal evidence, while non-causal proxy information is excluded from the mediator. The representation-level approximation of front-door adjustment can be written as:

$$p(Y \mid do(X)) \approx p(Y \mid M), \quad M = \mathrm{CCF}(Z_t^c, Z_v^c). \tag{3}$$

where Z_t^c and Z_v^c denote the causal sentiment evidence in text and image, respectively, and $CCF(\cdot)$ denotes the consistency-conflict aware fusion network. Meanwhile, the remaining modality-specific information in the inputs is separated into non-causal confounding proxies Z_t^n and Z_v^n.

3.3 CCF-SACE Framework

3.3.1 Sample-Adaptive Causal Evidence Selection

Given an image-text sample, we first extract token-level textual representations and patch-level visual representations:

$$\begin{aligned} H_t &= \mathrm{Enc}_t(X_t) \in \mathbb{R}^{L\times d_t}, \\ H_v &= \mathrm{Enc}_v(X_v) \in \mathbb{R}^{K\times d_v}. \end{aligned} \tag{4}$$

where L is the number of textual tokens, K is the number of image patches, $\mathrm{Enc}_t(\cdot)$ and $\mathrm{Enc}_v(\cdot)$ denote the text encoder and image encoder, respectively. Then, both modalities pass through two-layer adapters and are projected into a shared latent space, the average pooling is further applied to obtain global context representation for text and image:

$$\begin{aligned} \tilde{H}_t &= W_t H_t, \quad \tilde{H}_v = W_v H_v, \\ h_t &= \mathrm{Pool}(\tilde{H}_t), \quad h_v = \mathrm{Pool}(\tilde{H}_v). \end{aligned} \tag{5}$$

where W_t and W_v are learnable linear projection matrices. Different from selecting evidence within each modality independently, CCF-SACE estimates textual evidence under visual context and visual evidence under textual context. For the i-th textual token and the j-th image patch, the causal masks are computed as:

$$\begin{aligned} m_t^i &= \sigma\Big(\phi_t\Big([\tilde{h}_t^i; h_v; \tilde{h}_t^i \odot h_v; |\tilde{h}_t^i - h_v|]\Big)\Big), \\ m_v^j &= \sigma\Big(\phi_v\Big([\tilde{h}_v^j; h_t; \tilde{h}_v^j \odot h_t; |\tilde{h}_v^j - h_t|]\Big)\Big). \end{aligned} \tag{6}$$

Here, $m_t^i, m_v^j \in [0,1]$ are structural evidence selection weights, $\odot$ denotes element-wise multiplication, and $\phi_t(\cdot)$ and $\phi_v(\cdot)$ are learnable scoring networks. The masks jointly consider unimodal semantics, cross-modal agreement, and cross-modal discrepancy.

Based on the same masks, each modality is decomposed into a causal evidence stream and a non-causal proxy stream. For the textual modality, this decomposition is written as:

$$Z_t^c = \frac{\sum_{i=1}^{L} m_t^i \tilde{h}_t^i}{\sum_{i=1}^{L} m_t^i + \epsilon}, Z_t^n = \frac{\sum_{i=1}^{L} (1-m_t^i)\tilde{h}_t^i}{\sum_{i=1}^{L}(1-m_t^i) + \epsilon}. \tag{7}$$

The visual causal evidence Z_v^c and non-causal proxy Z_v^n are obtained analogously by replacing (t, i, L) with (v, j, K). The causal evidence stream is used to construct the mediator M, whereas the non-causal proxy stream is excluded from the main prediction path and used only for auxiliary suppression. We do not impose an additional mask-level loss. The selection masks are optimized implicitly through downstream prediction, causal evidence supervision, non-causal proxy suppression, and counterfactual sensitivity.

3.3.2 Consistency-Conflict Fusion

Textual and image causal evidence may exhibit agreement, complementarity, or conflict. To model their relation at the evidence level without returning to raw feature concatenation, we use a covariance-aware estimator to capture second-order dependencies between selected causal evidence units. Assume that Z_t^c and Z_v^c have already been processed in a normalized form. We compute the cross-modal covariance matrix and its compact relation summary as:

$$C_{tv} = \frac{1}{N-1}(Z_t^c - \mathbf{1}\mu_t^\top)^\top (Z_v^c - \mathbf{1}\mu_v^\top), \tag{8}$$

where μ_t and μ_v are the mean evidence vectors, $\mathbf{1}$ is an all-one vector. The covariance matrix is used to estimate dependency between selected causal evidence units, rather than to provide strict causal identification. The relation summary C_{tv} is then used as the agreement representation, and the scalar consistency score is obtained by:

$$M_{agr} = LN(C_{tv}), r = \sigma(w_r^\top C_{tv}), \tag{9}$$

where $r \in [0,1]$ measures the degree of cross-modal agreement, $\mathrm{LN}(\cdot)$ denotes layer normalization. A larger r indicates stronger agreement, while a smaller r suggests stronger discrepancy. Complementarity is not treated as simple inconsistency, it is handled by modality-specific conflict gates that preserve useful evidence from each modality.

For the conflict branch, we derive modality weights from the directional covariance strengths and directly fuse the two causal evidence vectors:

$$\begin{aligned} [\alpha_t, \alpha_v] &= \mathrm{softmax}\Big(\Big[\|C_{tv}\|_2, \|C_{tv}^\top\|_2\Big]/\tau\Big), \\ M_{con} &= \alpha_t Z_t^c + \alpha_v Z_v^c, \end{aligned} \tag{10}$$

where τ is a learnable temperature coefficient. The row and column covariance summaries respectively reflect how strongly textual and visual evidence are supported by the cross-modal relation. Finally, the evidence mediator is computed as:

$$M = rM_{agr} + (1-r)M_{con}. \tag{11}$$

When the two modalities are consistent, the model relies more on the covariance-derived agreement representation. When they are conflicting or weakly related, the model turns to the conflict branch and directly adjusts the contributions of textual and visual causal evidence.

3.4 Training Objectives

The training objectives are designed to support the evidence-mediated prediction path. First, the main task prediction is produced from the causal evidence mediator:

$$\begin{aligned} p_{main} &= \mathrm{softmax}(h_{main}(M)), \\ \mathcal{L}_{task} &= \mathrm{CE}(y, p_{main}). \end{aligned} \tag{12}$$

We impose direct supervision on causal evidence to ensure that the selected mediator retains sentiment-discriminative information:

$$p_{cb} = \mathrm{softmax}(h_{cb}([Z_t^c; Z_v^c])). \tag{13}$$

$$\mathcal{L}_{causal} = \mathrm{CE}(y, p_{cb}). \tag{14}$$

To prevent non-causal proxies from forming fast spurious prediction paths, constraining their prediction distribution to be close to the class-uniform distribution U is proposed:

$$\begin{aligned} p_{non} &= \mathrm{softmax}(h_{non}([Z_t^n; Z_v^n])), \\ \mathcal{L}_{non} &= D_{KL}(U \| p_{non}). \end{aligned} \tag{15}$$

Introducing a counterfactual sensitivity objective to make the evidence selector decision-sensitive rather than merely predictive is necessary. While $\mathcal{L}_{non}$ weakens the standalone predictability of non-causal proxies, $\mathcal{L}_{sens}$ further checks whether replacing the causal mediator with the proxy stream changes the model prediction. If p_{main} and the proxy-only prediction remain too similar, the selected evidence may not be sufficiently separated from shortcut information. We therefore construct a proxy-only counterfactual prediction and enforce a margin between the main and counterfactual predictions:

$$\begin{aligned} p_{cf} &= \mathrm{softmax}(h_{cf}([Z_t^n; Z_v^n])), \\ \mathcal{L}_{sens} &= \max(0, \gamma - D_{JS}(p_{main}, p_{cf})). \end{aligned} \tag{16}$$

Minimizing $\mathcal{L}_{sens}$ encourages the Jensen-Shannon divergence between p_{main} and p_{cf} to be at least γ, so that the main decision path relies on causal evidence rather than the proxy-only shortcut path. The final objective contains four functional losses:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_c \mathcal{L}_{causal} + \lambda_n \mathcal{L}_{non} + \lambda_s \mathcal{L}_{sens}. \tag{17}$$

During inference, we use p_{main} as the final sentiment prediction.

4 Experiments

We evaluate CCF-SACE on four MSA datasets and one fine-grained emotion recognition dataset, reporting Accuracy (Acc) and Weighted-F1 (F1). RoBERTa and ViT are used as the text and image encoders, with full experimental details provided in Appendix A.

4.1 Main Results

Table 1 compares CCF-SACE with representative unimodal and multimodal methods on five datasets. Overall, CCF-SACE achieves the best Acc and F1 on all datasets, showing that evidence-oriented fusion is effective across different multimodal sentiment and emotion recognition settings. It obtains 81.87% and 82.30% Acc on Twitter-15 and MVSA-Single, respectively, improving over the strongest baselines by 3.41 and 4.53 points. The improvements on Twitter-17 and MVSA-Multiple are smaller, 0.70 and 1.27 points in Acc, respectively. This may be because these datasets contain more ambiguous or weakly aligned image-text pairs, where sentiment labels depend on subtle annotation preferences rather than explicit causal evidence. These results indicate that CCF-SACE is not limited to one benchmark and can benefit datasets with different scales and sample distributions.

The gains on TumEmo are especially noteworthy. CCF-SACE improves Acc and F1 by 4.51 and 4.42 points over the strongest baseline. Since TumEmo is the largest dataset and contains seven fine-grained emotion categories, this result suggests that the proposed evidence-mediated fusion remains effective under large-scale and fine-grained affective supervision. We further report the multi-seed stability results on TumEmo in Appendix B.

4.2 Ablation studies

To verify the contribution of each component, we conduct ablations under the same RoBERTa+ViT backbone and training protocol. As shown in Table 2, removing cross-modal evidence selection, $\mathcal{L}_{causal}$, $\mathcal{L}_{non}$, or $\mathcal{L}_{sens}$ consistently degrades performance, indicating that evidence supervision, proxy suppression, and counterfactual sensitivity play complementary roles. The larger drops caused by removing CCF, applying CCF to raw features, or using fixed masks further support our central design choice: the model should first select sample-adaptive causal evidence and then perform consistency-conflict fusion at the evidence level.

Table 1: Experimental results (%) on multimodal sentiment and emotion datasets. Best results are in bold; unavailable results are denoted by "–".

Category	Model	TWITTER-15		TWITTER-17		MVSA-Single		MVSA-Multiple		TumEmo	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Text branch	BiLSTM (Wang and Yang, 2020)	–	–	–	–	70.12	65.06	67.90	67.90	61.88	51.26
	BERT (Devlin et al., 2019)	71.55	71.67	68.15	68.10	71.11	69.70	67.59	66.24	62.12	60.78
	RoBERTa (Liu et al., 2019)	73.02	72.93	69.66	69.66	71.17	70.74	67.51	66.46	62.38	61.43
Image branch	ResNet-50 (He et al., 2016)	52.84	50.70	46.60	45.18	64.67	61.55	61.88	60.98	46.35	45.94
	ViT (Dosovitskiy et al., 2021)	63.31	51.28	50.15	49.47	63.78	62.26	61.94	61.19	48.10	47.75
Multimodal	CLIP (Radford et al., 2021)	75.92	75.12	71.34	71.35	72.20	71.40	69.10	63.40	65.19	64.74
	MultiSentiNet (Xu and Mao, 2017)	–	–	–	–	69.84	69.63	68.16	68.11	63.09	53.98
	Se-MLNN (Cheema et al., 2021)	–	–	–	–	75.33	73.76	66.35	61.89	–	–
	CLMLF (Li et al., 2022)	–	–	–	–	75.33	73.46	72.00	69.83	68.10	68.00
	M2CL (Yang et al., 2023)	–	–	–	–	75.50	74.20	73.20	70.50	–	–
	TOM (Liu et al., 2024)	77.03	75.41	72.67	70.14	77.77	75.02	75.32	75.30	–	–
	CTMWA (Zhang et al., 2024)	75.24	74.95	71.36	70.87	75.91	75.74	74.02	73.84	68.57	68.60
	DRF (Wu et al., 2024)	75.03	75.41	69.42	68.52	74.50	74.40	71.00	68.20	69.60	69.60
	MultiAttention (Wang et al., 2025)	77.95	77.64	73.96	73.85	77.61	76.84	73.45	72.96	68.06	67.92
	CoDe (Wu et al., 2026)	78.46	78.12	74.58	74.43	76.98	76.47	72.74	70.71	73.02	73.02
Ours	**CCF-SACE**	**81.87**	**81.85**	**75.28**	**75.24**	**82.30**	**81.68**	**76.59**	**74.67**	**77.53**	**77.44**

Table 2: Ablation studies on MVSA-Single and TumEmo.

Variant	MVSA-Single		TumEmo	
	Acc	F1	Acc	F1
Full CCF-SACE	**82.30**	**81.68**	**77.53**	**77.44**
w/o evidence selection	80.12	79.36	76.38	76.25
$w/o\ \mathcal{L}_{causal}$	80.96	79.88	76.82	76.71
$w/o\ \mathcal{L}_{non}$	81.21	80.23	76.94	76.81
$w/o\ \mathcal{L}_{sens}$	80.74	79.45	76.71	76.52
w/o CCF	78.95	78.20	76.06	75.90
Raw-feature CCF	78.54	78.47	75.95	75.84
Fixed mask	79.08	78.12	76.12	75.98

Table 3: Same-backbone fusion comparison.

Fusion Strategy	MVSA-Single		TumEmo	
	Acc	F1	Acc	F1
Concatenation	76.55	74.99	75.57	75.50
Gated Fusion	76.88	75.55	75.68	75.60
Cross-Attention Fusion	77.21	76.04	75.81	75.72
Conflict-aware Fusion w/o Evidence Selection	78.54	78.47	76.36	76.24
CCF-SACE	**82.30**	**81.68**	**77.53**	**77.44**

To exclude the possibility that improvements mainly come from encoder capacity, Table 3 compares several fusion strategies under the same backbone. Simple concatenation and gated fusion provide limited gains, while cross-attention improves modality interaction but still operates on general fused representations. Conflict-aware fusion without evidence selection is stronger, yet remains below CCF-SACE. This comparison shows that relationship-aware fusion alone is insufficient; the key advantage comes from coupling relation modeling with sample-adaptive causal evidence selection.

Table 4: Shortcut robustness analysis.

Dataset	Model	Clean		Shortcut / Conflict		Drop
		Acc	F1	Acc	F1	Acc↓
MVSA -Single	Concatenation	76.55	74.99	68.74	66.80	7.81
	Attention Fusion	77.21	76.04	69.92	68.25	7.29
	CCF-SACE	**82.30**	**81.68**	**77.46**	**76.31**	**4.84**
TumEmo	Concatenation	75.57	75.50	68.82	68.41	6.75
	Attention Fusion	75.81	75.72	69.34	69.02	6.47
	CCF-SACE	**77.53**	**77.44**	**73.10**	**72.89**	**4.43**

4.3 Shortcut Robustness Analysis

Because CCF-SACE aims to reduce shortcut amplification, we evaluate it on shortcut-sensitive subsets, including textual shortcuts, visual shortcuts, and cross-modal conflicts. These subsets contain misleading sentiment words, spurious visual cues, or inconsistent text-image surface cues, testing whether the model still relies on shallow correlations under challenging distributions. Table 4 shows that all models degrade on shortcut-sensitive samples, but CCF-SACE yields the smallest Acc drop on both MVSA-Single and TumEmo. Its drop is 4.84 points on MVSA-Single, compared with 7.81 for concatenation and 7.29 for attention fusion; on TumEmo, its drop is 4.43 points, also lower than the two controlled baselines. This supports that selecting causal evidence before fusion reduces shortcut dependence rather than merely shifting it. Fine-grained results by shortcut type are reported in Appendix B.

4.4 Sensitivity and Evidence Diagnostics

We further analyze semantic evidence capacity and objective-weight sensitivity. Figure 4 varies the adaptor output dimension. Increasing the dimen-

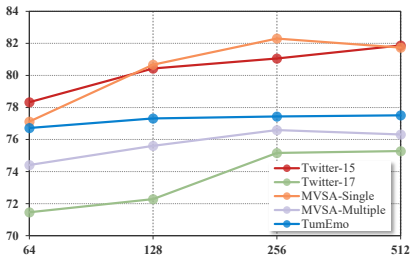

Figure 4: Effect of semantic evidence capacity.

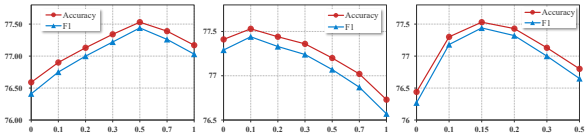

Figure 5: Sensitivity analysis of λ_c, λ_n, and λ_s.

sion from 64 to 256 generally improves performance, and larger benchmarks still benefit at 512 dimensions, suggesting that richer semantic evidence helps CCF-SACE construct more informative mediators. Figure 5 reports the sensitivity of three objective weights on TumEmo. The best performance appears around $\lambda_c = 0.5$, $\lambda_n = 0.1$, and $\lambda_s = 0.15$. Increasing λ_c improves performance by strengthening causal evidence supervision, whereas overly large non-causal suppression or sensitivity weights slightly degrade performance. This indicates that shortcut suppression should remain auxiliary rather than removing useful sentiment cues. Additional evidence-proxy diagnostics are provided in Appendix C.

4.5 Visualization

To qualitatively examine the learned representations, we visualize sentiment features on MVSA-Single, MVSA-Multiple, and TumEmo using t-SNE in Fig. 6. The representations learned by CCF-SACE form clearer and more compact clusters, which is consistent with the main results and suggests that evidence-mediated fusion helps separate affective categories.

Figure 7 presents two cases of causal visual evidence selection. The visualization highlights regions that contribute more to the causal evidence stream and contrasts them with non-causal regions. In both cases, the causal regions focus on visually prominent affective evidence and lead to reliable predictions, providing evidence-level interpretability beyond global fused features.

5 Conclusion

This work proposes CCF-SACE, a causal-inspired framework for MSA. We show that not all text

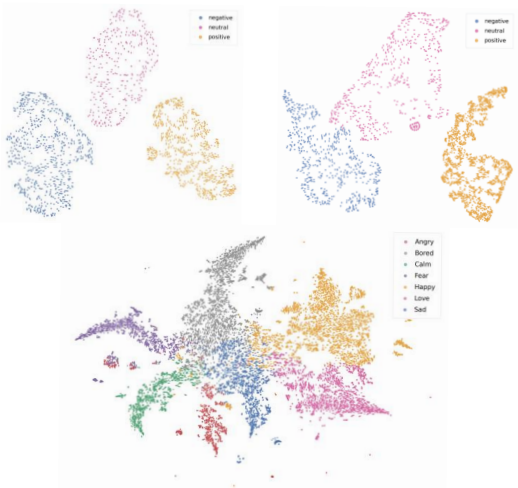

Figure 6: t-SNE visualization of sentiment representations on MVSA-Single, MVSA-Multiple, and TumEmo.

Figure 7: Case study of causal visual evidence selection.

tokens or image patches contribute equally to sentiment, and that the relationship between modalities can be consistent or conflicting. To address these issues, CCF-SACE learns cross-modal causal masks that decompose each modality into causal and non-causal evidence under the context of the other modality, and introduces a consistency-conflict aware fusion module that adaptively reweights evidence for agreement and conflict scenarios. In addition, counterfactual sensitivity objectives are imposed to ensure that causal evidence remains predictive while non-causal evidence becomes sentiment-agnostic. Comprehensive experiments on four multimodal sentiment analysis benchmarks and one multimodal emotion recognition dataset show that CCF-SACE achieves superior performance and robustness, offering a new perspective for causal-guided multimodal sentiment reasoning.

Limitations

While CCF-SACE performs well across multiple datasets, it still has some limitations that need to be addressed in future work. First, CCF-SACE uses learnable masks to separate causal sentiment evidence from non-causal interfering proxies. But the evidence selection is still guided by task supervision, so that some useful sentiment cues may not be fully selected, or some non-causal information may not be completely suppressed. Thus, future work will explore external knowledge or more explicit evidence supervision to improve the quality of causal evidence learning. Second, our work focuses mainly on image-text multimodal sentiment analysis. Images and texts are important for sentiment understanding. However, multimodal sentiment can also involve audio, video, and physiological signals. We plan to extend CCF-SACE to more diverse modality combinations and larger datasets. This will help evaluate the model's generalization ability. Furthermore, more challenging robustness settings, such as out-of-distribution, could provide a more complete understanding of the robustness of the model.

Ethical Considerations

This work studies multimodal sentiment analysis on publicly available social media datasets. The proposed framework does not introduce new data collection or annotation procedures and does not involve human subjects. Since social media datasets may contain annotation noise or dataset-specific biases, the results should be interpreted within the scope of the evaluated benchmarks.

References

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. 2021. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*, MMPT '21, page 37–45, New York, NY, USA. Association for Computing Machinery.

Fuhai Chen, Pengpeng Huang, Xuri Ge, Jie Huang, and Zishuo Bao. 2024. Multimodal sentiment analysis based on causal reasoning. *arXiv preprint arXiv:2412.07292*.

Jili Chen, Yihua Zhong, Qionghao Huang, Changqin Huang, Fan Jiang, Xiaodi Huang, and Xun Wang. 2025a. Ucmib-pns: balancing sufficiency and necessity with probabilistic causality and cross-modal uncertainty in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsen Zheng, and Liang Lin. 2025b. Cross-modal causal relation alignment for video question grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24087–24096.

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Xiaoran Hu and Masayuki Yamamura. 2022. Two-stage attention-based fusion neural network for image-text sentiment classification. In *Proceedings of the 2022 4th International Conference on Image, Video and Signal Processing*, pages 1–7.

Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. 2025. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725.

Shizhou Huang, Bo Xu, Changqun Li, Jiabo Ye, and Xin Lin. 2024. A sentimental prompt framework with visual text encoder for multimodal sentiment analysis. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 638–646.

Menghua Jiang, Yuxia Lin, Baoliang Chen, Haifeng Hu, Yuncheng Jiang, and Sijie Mai. 2025. Disentangling bias by modeling intra-and inter-modal causal attention for multimodal sentiment analysis. *arXiv preprint arXiv:2508.04999*.

Wenhui Li, Xinqi Su, Dan Song, Lanjun Wang, Kun Zhang, and An-An Liu. 2023. Towards deconfounded image-text matching with causal inference. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6264–6273.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2282–2294, Seattle, United States. Association for Computational Linguistics.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2149–2159. Association for Computational Linguistics.

Rui Liu, Jiahao Cao, Jiaqian Ren, Xu Bai, and Yanan Cao. 2025. Dual-path counterfactual integration for multimodal aspect-based sentiment classification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22759–22769.

Wuchao Liu, Wengen Li, Yu-Ping Ruan, Yulou Shu, Juntao Chen, Yina Li, Caili Yu, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. 2024. Weakly correlated multimodal sentiment analysis: New dataset and topic-oriented model. *IEEE Trans. Affect. Comput.*, 15(4):2070–2082.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jing Ma. 2025. Causal inference with large language model: A survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5886–5898.

Lei Meng, Xiangxian Li, Xiaoshuo Yan, Haokai Ma, Zhuang Qi, Wei Wu, and Xiangxu Meng. 2025. Causal inference over visual-semantic-aligned graph for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19449–19457.

Teng Niu, Shuhui Zhu, Liangliang Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multiview social data. In *MultiMedia Modeling (MMM 2016)*, volume 9517 of *Lecture Notes in Computer Science*, pages 15–27. Springer, Cham.

Ning Pang, Wansen Wu, Yue Hu, Kai Xu, Quanjun Yin, and Long Qin. 2024. Enhancing multimodal sentiment analysis via learning from large language model. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

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

Cheng Peng, Chunxia Zhang, Xiaojun Xue, Jiameng Gao, Hongjian Liang, and Zhengdong Niu. 2021. Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification. *Tsinghua Science and Technology*, 27(4):664–679.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Jing Ren, Wenhao Zhou, Bowen Li, Mujie Liu, Nguyen Linh Dan Le, Jiade Cen, Liping Chen, Ziqi Xu, Xiwei Xu, and Xiaodong Li. 2026. Causal prompting for implicit sentiment analysis with large language models. *IEEE Transactions on Computational Social Systems*.

Upendra Singh, Kumar Abhishek, and Hiteshwar Kumar Azad. 2024. A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, 56(9):1–38.

Teng Sun, Juntong Ni, Wenjie Wang, Liqiang Jing, Yinwei Wei, and Liqiang Nie. 2023. General debiasing for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5861–5869.

Hongbin Wang, Chun Ren, and Zhengtao Yu. 2025. Multimodal sentiment analysis based on multiple attention. *Eng. Appl. Artif. Intell.*, 140:109731.

Min Wang, Dong Cao, Lin Li, Shaozi Li, and Rongrong Ji. 2014. Microblog sentiment analysis based on cross-media bag-of-words model. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pages 76–80. ACM.

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, New York, NY, USA. Association for Computing Machinery.

Zihan Wang and Bo Yang. 2020. Attention-based bidirectional long short-term memory networks

for relation classification using knowledge distillation from BERT. In *IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress, DASC/PiCom/CBDCom/CyberSciTech 2020, Calgary, AB, Canada, August 17-22, 2020*, pages 562–568. IEEE.

Daiqing Wu, Dongbao Yang, Huawen Shen, Can Ma, and Yu Zhou. 2026. Resolving sentiment discrepancy for multimodal sentiment detection via semantics completion and decomposition. *Pattern Recognition*, 172:112719.

Daiqing Wu, Dongbao Yang, Yu Zhou, and Can Ma. 2024. Robust multimodal sentiment analysis of image-text pairs by distribution-based feature recovery and fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 5780–5789. ACM.

Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *2017 IEEE international conference on intelligence and security informatics (ISI)*, pages 152–154. IEEE.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 2399–2402. ACM.

Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. 2023. Multiple contrastive learning for multimodal sentiment analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021a. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021b. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9847–9857.

Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 189–198. ACM.

Baozheng Zhang, Ziqi Yuan, Hua Xu, and Kai Gao. 2024. Crossmodal translation based meta weight adaption for robust image-text sentiment analysis. *IEEE Trans. Multim.*, 26:9949–9961.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Danet: Dual-aware enhanced alignment network for multimodal aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14369–14381.

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. 2022. Multimodal sentiment analysis with image-text interaction network. *IEEE transactions on multimedia*, 25:3375–3385.

A Dataset and Implementation Details

We report the extended experimental setup omitted from Section 4. CCF-SACE is evaluated on Twitter-15, Twitter-17, MVSA-Single, MVSA-Multiple, and TumEmo. The first four datasets are three-class sentiment classification benchmarks with positive, neutral, and negative labels. Twitter-15 and Twitter-17 contain 3,502 and 2,910 multimodal tweet samples, respectively. MVSA-Single contains 4,869 paired image-text posts, while MVSA-Multiple contains 19,665 samples. TumEmo contains 195,265 samples, including 19,527 test samples, and covers seven emotion categories: Angry, Bored, Calm, Fear, Happy, Love, and Sad. This makes it a useful benchmark for testing whether the model learns scalable affective evidence rather than small-dataset artifacts.

All datasets use the same RoBERTa+ViT-based CCF-SACE framework. Twitter-15 and Twitter-17 follow the official train/dev/test splits, while MVSA-Single, MVSA-Multiple, and TumEmo use fixed 8:1:1 train/dev/test splits. The text and image encoders are kept consistent across datasets; only the dataset splits differ. Table 5 summarizes the reproducibility settings.

Table 5: Reproducibility details.

Setting	**Twitter**	**MVSA**	**TumEmo**
Text encoder checkpoint	RoBERTa-base		
Image encoder checkpoint	ViT-B/16, 224×224		
Maximum text length	192		
Image resolution	224×224		
Hidden dimension	256		
Batch size	64		
Optimizer	AdamW		
Learning rate	3×10^{-4}		
Weight decay	1×10^{-3}		
Warmup ratio	0.10		
Dropout	0.25		
λ_c	0.5		
λ_n	0.1		
λ_s	0.15		
γ	0.45		
Epochs / early stopping	40 / patience 8		

B Additional Stability and Robustness Results

Table 6 reports TumEmo results under five random seeds. The model achieves 77.48 ± 0.06 Acc and 77.36 ± 0.09 F1. The small standard deviations indicate that the improvement on TumEmo is reproducible rather than caused by a favorable random initialization. Together with Table 1, these results show that the gain on TumEmo is both strong and stable.

Table 6: TumEmo multi-seed results.

Seed	Acc	F1
101	77.53	77.44
102	77.38	77.22
103	77.49	77.35
104	77.47	77.42
105	77.53	77.39
Mean ± Std	77.48 ± 0.06	77.36 ± 0.09

Table 7 further breaks down shortcut robustness by shortcut type. CCF-SACE improves over the baseline on textual shortcuts, visual shortcuts, cross-modal conflict, and weakly related image-text pairs. The largest gap appears on cross-modal conflict samples, where naive fusion is more likely to amplify inconsistent surface cues. The improvement on weakly related image-text pairs also suggests that the model can avoid overusing a modality when it provides limited sentiment support.

Table 7: Fine-grained shortcut subset results.

Subset	**Baseline**		**CCF-SACE**	
	Acc	F1	Acc	F1
Textual shortcut	70.42	69.84	76.08	75.76
Visual shortcut	69.15	68.52	74.96	74.50
Cross-modal conflict	66.87	65.91	73.48	72.86
Weakly related image-text	71.34	70.62	75.02	74.61

C Confusion Matrix and Evidence Diagnostics

We provide confusion matrices on all five datasets in Fig. 8. CCF-SACE generally produces strong diagonal patterns, indicating that the learned causal evidence mediator preserves class-discriminative sentiment information. On Twitter-15 and Twitter-17, most errors occur between positive and neutral or negative and neutral classes, suggesting that neutral expressions are ambiguous in short social media posts. On MVSA-Single and MVSA-Multiple, positive samples are recognized with high accuracy, while neutral samples are more frequently confused with positive ones, reflecting subjectivity and annotation ambiguity in image-text sentiment labels. On TumEmo, the model recognizes Love, Happy, Bored, and Calm relatively well, whereas Angry and Sad show more confusion with adjacent emotions. These observations are consistent with

Figure 8: Confusion matrices of CCF-SACE on five multimodal sentiment and emotion benchmarks.

Table 8: Evidence-proxy separation diagnostics.

Model	Proxy Ent. ↑	D_{JS} ↑	Mask Conc.	Text Ratio	Visual Ratio
w/o $\mathcal{L}_{non}$	0.63	0.18	1.71	0.43	0.39
w/o $\mathcal{L}_{sens}$	0.82	0.11	1.58	0.46	0.42
CCF-SACE	**0.91**	**0.27**	2.34	0.31	0.28

the main results and further show that CCF-SACE is effective on both coarse-grained sentiment classification and fine-grained emotion recognition.

Table 8 reports quantitative diagnostics for evidence-proxy separation. A desirable model should keep the proxy-only prediction less confident, maintain sufficient divergence between the main and proxy-only predictions, and select compact but not degenerate evidence regions. Full CCF-SACE achieves the highest proxy entropy and JS divergence, indicating that the proxy stream becomes less independently predictive while the main and proxy-only paths are more clearly separated. Its higher mask concentration and lower evidence ratios also suggest that the model selects compact evidence instead of spreading attention over potentially spurious regions.

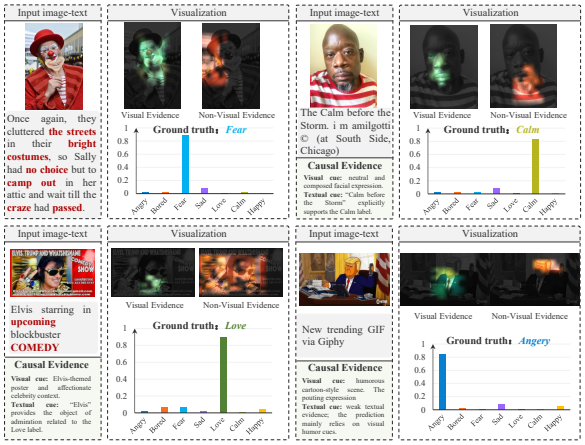

Figure 9: Additional case studies of causal visual evidence selection.

D Additional Case Studies

Figure 9 shows additional causal visual evidence selection cases. The visualization highlights image regions that contribute more strongly to the causal evidence stream and contrasts them with non-causal regions. These examples further support the evidence-level interpretability of CCF-SACE across diverse affective categories.