CVPR
#18405

TCMR: Text Confidence-aware Missing Semantic Reconstruction for Incomplete Multimodal Sentiment Analysis

Anonymous CVPR submission

Paper ID 18405

Abstract

*Recent advances in multimodal sentiment analysis (MSA) have predominantly adopted text-guided fusion, leveraging the semantic richness of language to integrate cues from other modalities. However, performance often degrades in realistic settings with missing modalities, particularly when the text modality is incomplete. This paper presents **T**ext **C**onfidence-aware **M**issing **R**econstruction (**TCMR**), a framework that reconstructs missing semantics in incomplete text using confidence scores that quantify textual informativeness. TCMR first generates pseudo confidence labels using a pretrained sentiment-polarity classifier, and the Text Confidence Estimation (TCE) module learns to predict these scores from incomplete text. In parallel, an Importance-aware Proxy Feature Generator (IPFG) produces proxy text features from auxiliary modalities, adaptively weighting them by their contribution to reconstruction. The predicted confidence then weights the combination of incomplete text features and proxy features, enabling confidence-guided reconstruction. To stabilize joint training, where the confidence estimator shares representations with the main network, we introduce an Alternative Optimization Strategy (AOS) that balances the two objectives. Experiments on MOSI and MOSEI datasets demonstrate that TCMR reconstructs semantically meaningful text representations and outperforms prior reconstruction-based methods under missing-modality conditions.*

1. Introduction

Traditional sentiment analysis [12, 13] focuses on inferring human emotions from linguistic cues. The growth of social media platforms such as YouTube and TikTok has enabled the collection of large-scale multimodal data, making Multimodal Sentiment Analysis (MSA) an active research field that jointly analyzes nonverbal (e.g., facial expressions) and paralinguistic cues (e.g., tone) alongside text. Building on the rich sentiment-relevant information in text [8, 20], recent MSA studies have increasingly explored text-centric approaches [5, 18, 19, 26] that treat text as the dominant modality while considering audio and vision modalities as auxiliary sources. Although these approaches achieve strong performance with fully observed training and test data, they often struggle to generalize to real-world scenarios where modalities are partially missing, known as the missing modality problem.

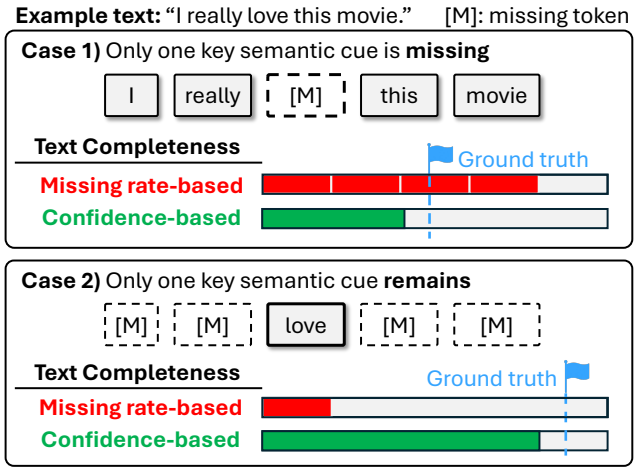


Figure 1. Comparison between missing rate–based and confidence–based completeness. Note that the former is computed as the complement of the missing rate, while the latter is estimated by our proposed method.

To mitigate this problem, several studies have proposed reconstruction-based approaches [23, 27, 28], aiming to restore missing semantics. Among them, [27] introduces a completeness-based reconstruction framework called LNLN, which reconstructs the corrupted text modality by generating proxy features from auxiliary modalities and fusing them according to the missing rate. However, completeness estimation based on the missing rate can misrepresent semantic informativeness, leading to suboptimal reconstruction results. As illustrated in Figure 1, two contrasting cases highlight this problem: In Case 1, although

CVPR #18405

the missing rate is low, the absence of a key sentiment token leads to substantial semantic loss, resulting in an overestimation of completeness. In Case 2, the overall sentiment meaning is preserved even with a high missing rate but completeness is underestimated.

Inspired by the above observations, we propose the **T**ext **C**onfidence-aware **M**issing **R**econstruction (**TCMR**) framework that robustly estimates the semantic informativeness in the corrupted text, thereby enabling more accurate reconstruction under arbitrary missing conditions. Specifically, in the Text Confidence Estimation (TCE) module, TCMR estimates a confidence score of the incomplete text feature to assess the extent to which sentiment-relevant semantic information is preserved. Meanwhile, the Importance-aware Proxy Feature Generator (IPFG) generates proxy features from the auxiliary modalities by adaptively weighting their contributions with respect to text reconstruction. Subsequently, TCMR reconstructs the incomplete text features by weighting the proxy features with the estimated confidence score. Then, text-guided multimodal fusion is conducted with the reconstructed text features as the dominant modality for predicting the final sentiment score. Furthermore, an Alternative Optimization Strategy (AOS) is applied to stabilize TCMR training by decoupling confidence learning from the other objectives to prevent interference between optimization signals.

In summary, this work makes three main contributions. First, we reveal the limits of missing rate–based importance estimation in capturing sentiment-relevant semantics and introduce a semantics-aware Textual Completeness Estimation (TCE) module that directly models textual informativeness. Second, to further enhance TCE and stabilize the training process, we introduce an Importance-aware Proxy Feature Generation (IPFG) module and an Alternative Optimization Strategy (AOS), respectively. Third, through extensive experiments against nine competitive baselines, we demonstrate that the proposed TCMR framework consistently outperforms existing methods, achieving gains of 2.29 Non0 Acc points on MOSI and 0.46 Has0 Acc points on MOSEI.

2. Related Work

Recent studies in MSA can be broadly categorized by their fusion paradigms: ternary-symmetric and text-centric. Ternary-symmetric approaches [7, 9, 15, 16, 22, 25] treat all modalities with equal importance and focus on learning joint representations that capture inter-modal correlations. For example, Self-MM [22] jointly models modality consistency and specificity through self-supervised objectives, and MMIM [7] maximizes mutual information between unimodal inputs and fused representations to preserve modality-specific information. By contrast, text-centric approaches [5, 18, 19, 26] view text as the dominant modality, leveraging audio and vision signals as auxiliary cues to enrich textual semantics. CENet [18] enhances textual representations by injecting emotion-related features from auxiliary modalities, while ALMT [19] filters out sentiment-irrelevant noise guided by textual context. These methods achieve strong performance when all modalities are available but often suffer degradation under missing inputs.

To address such real-world incompleteness, recent studies have explored reconstruction-based techniques that aim to recover missing modality features. TFR-Net [23] employs a transformer-based architecture to reconstruct randomly dropped features across modalities, and LNLN [27] adopts a text-centric design to restore missing textual semantics using proxy features derived from audio and vision inputs, where the fusion is weighted by the estimated missing rate to approximate textual completeness. More recently, P-RMF [29] introduces a proxy-driven strategy that dynamically reconstructs incomplete multimodal inputs through cross-modal feature generation and adaptive fusion, further improving robustness under uncertain missing conditions. While these methods enhance resilience to missing modalities, they primarily rely on structural or rate-based reconstruction objectives and thus overlook semantic informativeness in text.

3. Proposed Framework

3.1. Problem Definition

Multimodal sentiment analysis is formulated as a regression problem that predicts a sentiment score from synchronized text (t), audio (a), and vision (v) inputs derived from the same utterance. In this study, we consider a random missing scenario where each modality is partially missing within the sequence while calling for robust models that remain reliable under incomplete observations.

3.2. Framework Overview

The proposed framework TCMR is illustrated in Figure 2. It is composed of three main parts: a modality-specific encoder that transforms each modal input into a feature representation (①); a semantic module that estimates text confidence based on the text semantics and reconstructs text semantics from the auxiliary modalities (②–④); and a text-centric fusion module that combines the reconstructed text features with the auxiliary features to produce the final sentiment prediction (⑤). We describe modules focused on their objectives in the main paper and elaborate on detailed network architectures in Table A.

3.3. Input Processing

Raw multimodal inputs are processed into high level embeddings (Figure 2 ①). We denote the complete raw multimodal inputs as I_m^c for each modality $m \in \{t, a, v\}$. For

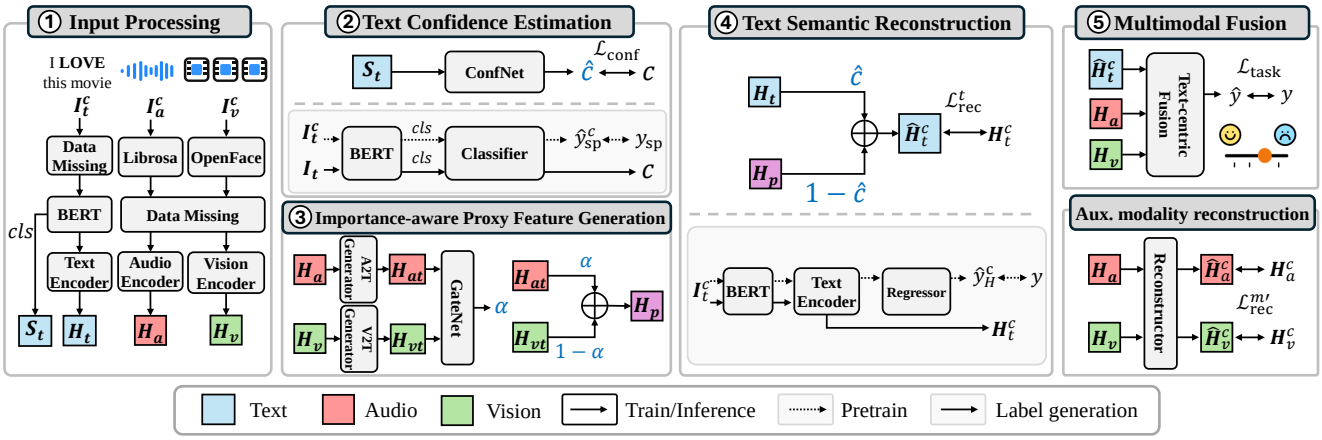


Figure 2. Overview of the proposed TCMR framework.

the complete text I_t^c, we randomly replace 0 - 100% of the tokens with `[UNK]` following prior works [27, 28], and obtain incomplete text $I_t = \text{DataMissing}_t(I_t^c)$. Then we obtain low level feature X_t by utilizing BERT encoder [4] parameterized by θ^{bert}:

$$X_t = \text{BERT}(I_t; \theta^{\text{bert}}). \tag{1}$$

For audio I_a^c and vision I_v^c, we utilize Librosa [11] and OpenFace [2] and replace 0 - 100% of temporal segments with zero vectors following the same prior works [27, 28], yielding low level feature X_a and X_v:

$$X_a = \text{DataMissing}_a(\text{Librosa}(I_a^c)), \tag{2}$$
$$X_v = \text{DataMissing}_v(\text{OpenFace}(I_v^c)). \tag{3}$$

Each low level feature $X_m \in \mathbb{R}^{T_m \times d_m}$ has temporal dimension of T_m and feature dimension of d_m.

Subsequently, each X_m is processed by a modality-specific Transformer encoder θ_m^{trans} [17] to obtain high-level representation $H_m \in \mathbb{R}^{T \times d}$:

$$H_m = \text{Transformer}([X_m, E_m]; \theta_m^{trans}), \tag{4}$$

where $E_m \in \mathbb{R}^{T \times d}$ denotes a learnable embedding initialized with a token length of T and [,] indicates the concatenation operation.

3.4. Confidence-guided Text Reconstruction

Text Confidence Estimation (TCE) A confidence estimator `ConfNet`, which is the central contribution of this work, quantifies the semantic informativeness of a text I_t and enables informativeness-aware multimodal fusion for robust restoration of missing textual information (Figure 2 ②). Given the BERT `[CLS]` embedding S_t, `ConfNet` estimates a scalar confidence $\hat{c} \in [0, 1]$:

$$\hat{c} = \text{ConfNet}(S_t; \theta^{\text{conf}}), \tag{5}$$

where θ^{conf} are model parameters.

Specifically, `ConfNet` is a multilayer perceptron composed of fully connected layers followed by a sigmoid activation. θ^{conf} is optimized to minimize the following loss that quantifies mean squared error between the estimated $\hat{c}$ and the pseudo confidence label c:

$$\mathcal{L}_{\text{conf}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{c}^{(i)} - c^{(i)} \right\|^2, \tag{6}$$

where i denotes the data sample index. The resulting θ^{conf} acts as a reliability signal that balances text and other features in Text Semantic Reconstruction step (Figure 2 ④).

To optimize `ConfNet` with Equation 6, we first prepare a pseudo confidence label c prior to the main training phase (Figure 2 ②, grey section). To that end, we adopt the True Class Probability criterion [3], a measure of classification reliability that uses the probability assigned to the ground truth class as the reliability. We train a three class sentiment polarity classifier with softmax output that predict sentiment in positive, neutral, and negative on complete text I_t^c (dotted line). Then we compute the pseudo confidence c as the probability that the classifier assigns to the correct polarity class (solid line) given incomplete text I_t:

$$c^{(i)} \triangleq p\left(y_{\text{sp}}^{(i)} \mid I_t^{(i)}; \theta_{\text{pre}}^{\text{classifier}}\right), \tag{7}$$

where $y_{\text{sp}}^{(i)}$ denotes the ground truth sentiment label and $\theta_{\text{pre}}^{\text{classifier}}$ are the classifier parameters. Intuitively, if semantic cues present in the complete text are missing in the incomplete input, the probability for the correct class naturally decreases, which makes c a principled supervisory target for learning `ConfNet`. The validity of this pseudo confidence design was verified through the sentiment clue sensitivity analysis described in Section 5.

CVPR #18405

Importance-aware Proxy Feature Generation (IPFG) To convert $H_{m' \in \{a,v\}}$ to features that complement text feature H_t, we introduce A2TGenerator and V2TGenerator. (Figure 2 ③):

$$\begin{aligned} H_{at} &= \text{A2TGenerator}([H_a,\, E_a^g];\, \theta_a^{\text{gen}})\,, \\ H_{vt} &= \text{V2TGenerator}([H_v,\, E_v^g];\, \theta_v^{\text{gen}})\,. \end{aligned} \tag{8}$$

Each generator takes feature $H_{m'}$ together with a randomly initialized learnable embedding $E_{m'}^g$ and produces a representation $H_{m't}$.

The amount of semantic cue present in H_{vt} and H_{at} can differ depending on their contents and missing rate. We therefore employ a gating network that outputs a scalar weight $\alpha \in [0, 1]$, and the proxy representation H_p is obtained by a weighted sum:

$$\begin{aligned} \alpha &= \text{GateNet}\big([H_{at},\, H_{vt}];\, \theta^{\text{gate}}\big)\,, \\ H_p &= \alpha\, H_{at} \;+\; (1-\alpha)\, H_{vt}. \end{aligned} \tag{9}$$

This gating allows the model to emphasize the modality $m' \in \{a, v\}$ that carries stronger semantic cues for the current utterance while down weighting the other.

Text Semantic Reconstruction Finally, we reconstruct the complete text representation $\hat{H}_t^c \in \mathbb{R}^{T \times d}$ by integrating the incomplete text feature H_t and the generated proxy feature H_p based on the predicted confidence score c from `ConfNet`:

$$\hat{H}_t^c = cH_t + (1-c)H_p. \tag{10}$$

To ensure the reconstructed $\hat{H}_t^c$ matches the feature of complete text H_t^c, we minimize the following loss (Figure 2 ④):

$$\mathcal{L}_{\text{rec}}^t = \frac{1}{N}\sum_{i=1}^{N} \left\| \hat{H}_t^{c(i)} - H_t^{c(i)} \right\|^2. \tag{11}$$

Specifically, $H_t^{c(i)}$ is prepared before the main training from a sentiment regression model (Figure 2 ④, grey section). The regression model is optimized to minimize mean squared error between predicted $\hat{y}_H^c$ and true sentiment y (dotted line). After the training, the output representation H_t^c of Text Encoder is used as a complete text feature for a corresponding incomplete text feature H_t.

3.5. Multimodal Fusion and Sentiment Prediction

The multimodal fusion module follows the design of a prior work [27], and its key components are briefly summarized here. Starting from the reconstructed text representation H_p, a refinement Transformer encoder applies self attention, producing layerwise refined features $H_t^{+(i)}$ with $i \in \{1, ..., L_{\text{ref}}\}$. Based on $H_t^{+(i)}$, cross modal attention integrates complementary cues from vision and audio across multiple layers: at each layer, the current refined text feature $H_t^{+(i)}$ serves as the query and attends to H_v and H_a, while the fused representation updates $H_f^{(i)}$ by residual accumulation; in compact form we write $H_f^{(i)} = H_f^{(i-1)} + \text{MHA}(H_t^{+(i)}, H_v) + \text{MHA}(H_t^{+(i)}, H_a)$, where H_f^0 is a learnable embedding, and $\text{MHA}(Q, K)$ denotes multi head attention with query Q and key and value from K. A CrossTransformer then models interactions between $[\,H_f^{(L_{\text{ref}})},\, H_t^{+(L_{\text{ref}})}\,]$ and a regression head outputs the final sentiment prediction $\hat{y} = g_{\text{cross}}([\,H_f^{(L_{\text{ref}})},\, H_t^{+(L_{\text{ref}})}\,];\, \theta^{\text{cross}})$; training minimizes mean squared error:

$$\mathcal{L}_{\text{task}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{y}^{(i)} - y^{(i)}\|^2. \tag{12}$$

3.6. Auxiliary Modality Reconstruction

A Reconstructor model [27] is employed to encourage to encourage the audio encoder θ_a^{trans} and vision encoder θ_v^{trans} to embed as much information as possible (Figure 2, Aux. modality reconstruction). Given an incomplete feature $H_{m'}$ with $m' \in \{a, v\}$, the Reconstructor outputs a prediction $\hat{H}_{m'}^c = \text{Transformer}(H_{m'}; \theta_{m'}^{recon})$ of the corresponding complete feature $H_{m'}^c$. Reconstructor is optimized to minimize the discrepancy between $\hat{H}_{m'}^c$ and $H_{m'}^c$, while stop gradient is applied to $H_{m'}^c$:

$$\mathcal{L}_{\text{rec}}^{m'} = \frac{1}{N}\sum_{m' \in \{a,v\}}\sum_{i=1}^{N} \left\| \hat{H}_{m'}^{(i)} - H_{m'}^{\text{c}(i)} \right\|^2. \tag{13}$$

This objective jointly trains the audio and vision encoders so that $H_{m'}$ retains necessary information to reconstruct the original complete inputs, which in turn improves the final performance.

3.7. Alternative Optimization Strategy for Stable Learning

The final training objective is a weighted sum of module-wise losses:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{conf}} + \beta\mathcal{L}_{\text{rec}}^t + \gamma\mathcal{L}_{\text{rec}}^{m'} + \sigma\mathcal{L}_{\text{task}}. \tag{14}$$

We found naively minimizing $\mathcal{L}_{\text{total}}$ in an end to end manner often resulted in unstable training, where $\mathcal{L}_{\text{conf}}$ was effectively ignored and `ConfNet` remains at a poor local minimum.

We attribute this behavior to gradient conflict [21]. Viewed over long optimization process, the two objectives $\mathcal{L}_{\text{conf}}$ and $\mathcal{L}_{\text{task}}$ are aligned toward the same end goal. The confidence loss $\mathcal{L}_{\text{conf}}$ trains `ConfNet` so that reconstruction and fusion allocate emphasis to informative text, which in expectation reduces $\mathcal{L}_{\text{task}}$. However, at the scale of a single optimization step, the situation can differ. The gradient

Algorithm 1 Alternative Optimization Strategy

Input: dataset $\mathcal{D}$, epochs E, batch size B, weights β, γ.
Params: $\Theta^{\text{conf}} = \theta^{\text{conf}} \cup \theta^{\text{bert}}$, Θ^{other} for the remaining modules including θ^{bert}.
Opt: Opt_{conf}, $\text{Opt}_{\text{other}}$.

for $e = 1$ to E **do**
  ▷ Phase A. optimize Θ^{conf} to minimize $\mathcal{L}_{\text{conf}}$.
  for mini-batch $\mathcal{B} \subset \mathcal{D}$, $|\mathcal{B}| = B$ **do**
    $\mathcal{L}_{\text{conf}} \leftarrow \text{LossConf}(\mathcal{B}; \Theta^{\text{conf}})$
    $\text{Opt}_{\text{conf}}.\text{step}(\nabla_{\theta^{\text{conf}}} \mathcal{L}_{\text{conf}})$
  end for
  ▷ Phase B. optimize Θ^{other} to minimize $\mathcal{L}_{\text{other}}$.
  for mini-batch $\mathcal{B} \subset \mathcal{D}$, $|\mathcal{B}| = B$ **do**
    $\mathcal{L}_{\text{task}} \leftarrow \text{LossTask}(\mathcal{B}; \Theta^{\text{other}})$
    $\mathcal{L}^{t}_{\text{rec}} \leftarrow \text{LossRecText}(\mathcal{B}; \Theta^{\text{other}})$
    $\mathcal{L}^{m'}_{\text{rec}}, \leftarrow \text{LossRecAux}(\mathcal{B}; \Theta^{\text{other}})$
    $\mathcal{L}_{\text{other}} \leftarrow \mathcal{L}_{\text{task}} + \beta\, \mathcal{L}^{t}_{\text{rec}} + \gamma\, \mathcal{L}^{m'}_{\text{rec}}$
    $\text{Opt}_{\text{other}}.\text{step}(\nabla_{\Theta^{\text{other}}} \mathcal{L}_{\text{other}})$
  end for
end for

induced by $\mathcal{L}_{\text{conf}}$ on the shared parameters θ^{conf} and θ^{bert} does not always imply a decrease of $\mathcal{L}_{\text{task}}$. Formally, the cosine between $\nabla_{\{\theta^{\text{conf}}, \theta^{\text{bert}}\}} \mathcal{L}_{\text{conf}}$ and $\nabla_{\{\theta^{\text{conf}}, \theta^{\text{bert}}\}} \mathcal{L}_{\text{task}}$ can be negative, that is

$$\langle \nabla \mathcal{L}_{\text{conf}},\, \nabla \mathcal{L}_{\text{task}} \rangle < 0,$$

which indicates a step level gradient conflict.

To mitigate this issue, we propose an Alternative Optimization Strategy (AOS), Algorithm 1) that partitions the objective into

$$\mathcal{L}_{\text{conf}} \quad \text{and} \quad \mathcal{L}_{\text{other}} = \beta\, \mathcal{L}^{t}_{\text{rec}} + \gamma\, \mathcal{L}^{m'}_{\text{rec}} + \sigma \mathcal{L}_{\text{task}}. \tag{15}$$

Within each training epoch, we perform two consecutive phases. First, we optimize `ConfNet` θ^{conf} and preceding θ^{bert} by minimizing $\mathcal{L}_{\text{conf}}$ while keeping the remaining modules fixed. Second, we optimize the rest of the model by minimizing $\mathcal{L}_{\text{other}}$.

4. Experiments

4.1. Experimental Setup

Dataset We conduct experiments on two MSA benchmark datasets, MOSI [24] and MOSEI [1]. MOSI consists of 2,199 utterance-level samples, which are split into 1,284 for training, 229 for validation, and 686 for testing. MOSEI contains a total of 22,856 samples, with 16,326 for training, 1,871 for validation, and 4,659 for testing. In both datasets, sentiment scores are labeled by human annotators with continuous values ranging from –3 (strongly negative) to +3 (strongly positive).

Table 1. Hyperparameters of TCMR used for the MOSI and MOSEI datasets

	MOSI	MOSEI
Learning Rate	1e-4	1e-4
Weight Decay	1e-4	1e-4
Epochs	200	200
Batch Size	64	32
Loss Weight $\alpha, \beta, \gamma, \sigma$	0.1, 0.8, 0.1, 1.0	0.1, 0.8, 0.1, 1.0
Warm up	✓	✓
AOS	✓	✓
Early Stop	✓	✓
Patience	10	10
Seed	1111,1112,1113	1111,1112,1113

Evaluation Metrics To ensure generalizability, we train the model with three different random seeds and report the average performance on the test set. To comprehensively assess model performance, we use six different metrics following prior work [27, 28]: classification accuracy for 7-class and 5-class sentiment prediction (Acc-7 and Acc-5), binary classification accuracy and F1-score (Acc-2 and F1), and mean absolute error (MAE) and correlation (Corr) for regression performance.

Training and Evaluation Settings Following previous studies [27, 28], we adopt a *partial random missing* scenario. During training, we randomly erase each modality with a missing rate r sampled from a uniform distribution in the range of $[0, 1.0]$. The best model on the validation dataset is selected at $r = 0.5$. For testing, we conduct experiments ten times, setting the missing rate r from 0 to 0.9 with an increment of 0.1. The detailed hyperparameter configurations are provided in Table 1

4.2. Main Experimental Results

Baseline Models In the experiments, we compare the proposed TCMR framework with two groups of existing MSA approaches. *Non-reconstructing methods* include MISA [7], Self-MM [22], MMIM [6], CENet [18], TETFN [19], and ALMT [26], which do not reconstruct missing modalities. *Reconstruction-based methods* include TFR-Net [23], LNLN [27], and P-RMF [28].

For a fair comparison, we reproduce all baseline models using their official implementations under the same experimental environment as TCMR. Specifically, MISA, Self-MM, MMIM, CENet, TETFN, TFR-Net, and ALMT are reproduced based on the open-source code provided in the MMSA [10]. LNLN[1] and P-RMF[2] are reproduced from their official GitHub repositories.

[1] https://github.com/Haoyu-ha/LNLN
[2] https://github.com/aoqzhu/P-RMF

Table 2. Robustness comparison of the overall performance on MOSI and MOSEI datasets. Note: The smaller MAE indicates the better performance.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Non0 Acc / F1	Has0 Acc / F1	MAE	Corr	Acc-7	Acc-5	Non0 Acc / F1	Has0 Acc / F1	MAE	Corr
MISA	29.03	31.61	68.77 / 68.67	67.94 / 67.72	1.1637	47.57	43.89	44.43	72.22 / 68.21	74.16 / 71.24	0.7346	43.57
Self-MM	30.38	33.69	68.83 / 68.63	68.65 / 69.34	1.1843	47.25	46.45	47.36	73.02 / 71.43	72.66 / 71.82	0.6818	53.68
MMIM	30.67	34.31	70.19 / 69.87	69.59 / 69.15	1.1649	49.01	44.89	45.42	74.29 / 73.19	73.46 / 73.19	0.7032	52.75
CENet	29.49	32.90	69.88 / 69.94	69.43 / 69.39	1.1801	48.24	**47.36**	**48.24**	77.01 / 77.26	75.96 / 76.23	0.6622	58.24
TETFN	29.86	32.56	70.62 / 70.67	69.85 / 69.79	1.1327	49.16	46.78	47.83	77.87 / 77.45	76.11 / 76.37	0.6741	58.31
TFR-Net	28.22	30.31	70.88 / 70.73	70.18 / 69.93	1.1454	50.05	46.08	46.47	75.46 / 74.10	74.18 / 73.61	0.6784	56.24
ALMT	29.57	32.05	71.51 / 71.48	70.52 / 70.39	1.1671	47.57	46.66	47.37	76.83 / 76.41	74.47 / 74.84	0.6749	56.45
LNLN	31.36	34.43	70.00 / 69.99	69.51 / 69.41	1.1450	48.08	46.36	47.14	77.79 / 77.27	76.43 / 76.58	0.6698	58.17
P-RMF	28.43	30.01	69.39 / 70.06	68.73 / 69.53	1.1361	48.27	45.29	46.18	**78.46 / 77.76**	77.13 / 77.11	0.6738	58.62
TCMR	**32.79**	**36.72**	**72.01 / 71.66**	**70.93 / 70.49**	**1.0720**	**52.28**	47.27	48.16	77.99 / 76.99	**77.59 / 77.21**	**0.6614**	**58.63**

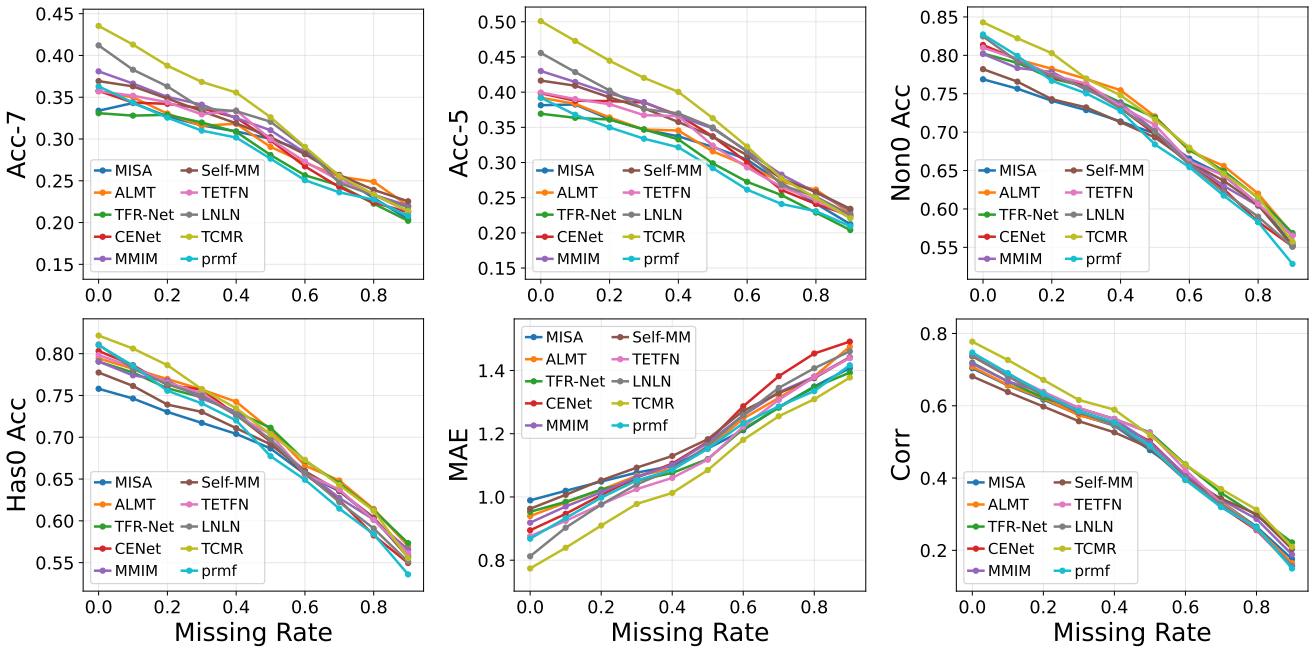
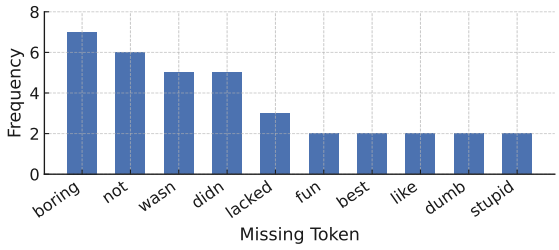


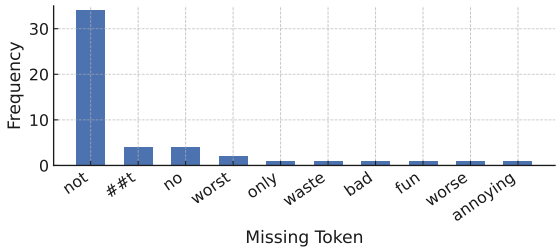
Figure 3. Comparison of model performance under different missing rates on MOSI.

Overall Performance Table 2 presents the evaluation results across various models on the MOSI and MOSEI datasets. On the MOSI dataset, our proposed TCMR achieves state-of-the-art performance across all metrics, demonstrating remarkable robustness under arbitrary missing conditions. In particular, TCMR improves Acc-5 by 6.7% and reduces MAE by 6.4% compared to LNLN on the MOSI dataset, verifying the superiority of our confidence-based completeness in reconstructing incomplete text. To examine model robustness in detail, we conduct a fine-grained analysis using performance curves under different missing rates. As illustrated in Figure 3, TCMR consistently maintains higher performance than other baselines across varying missing rates. These results indicate that TCMR adaptively adjusts confidence across varying missing rates, thereby maximizing the effectiveness of multimodal fusion and ensuring reliable applicability in diverse real-world scenarios with modality incompleteness.

On the MOSEI dataset, TCMR achieves state-of-the-art performance on Has0 Acc / F1, MAE, and Corr, while remaining competitive on Acc-7, Acc-5, and Non0 Acc / F1. One noteworthy observation is that several non-reconstructing baselines also exhibit relatively strong performance on MOSEI, consistent with prior observa-



(a) MOSI



(b) MOSEI

Figure 4. Comparison of error-triggering sentiment clue tokens across two benchmark datasets. Each histogram shows the frequency of missing tokens that cause misclassification in the sentiment classifier.

tions [27]. This phenomenon is largely attributed to a *lazy behavior* under class imbalance, where the model becomes biased toward the majority class. Since the neutral class dominates both the training and testing sets in MOSEI, these models tend to over-predict the neutral class, inflating aggregate metrics without faithfully modeling fine-grained sentiment.

5. In-depth Analysis

Sentiment Clue Sensitivity Analysis To validate whether the text confidence estimator `ConfNet` effectively captures semantic informativeness, an experiment was conducted to assess its sensitivity to sentiment-bearing tokens. A sentiment classifier was trained on complete textual inputs, and only correctly predicted samples were used for analysis. Each sample was perturbed by masking one token at a time, and the number of cases in which this perturbation turned a previously correct prediction into an incorrect one was counted for each token, as visualized in Figure 4. Misclassifications were concentrated on affective words such as *love*, *hate*, and *boring*. This pattern demonstrates that the sentiment classifier relies on these key tokens as primary affective cues, and that `ConfNet` is accurately supervised to capture their semantic importance by the pseudo confidence labels, leading to a reliable and semantically grounded confidence signal for multimodal fusion.

Qualitative comparison A closer look at each quadrant in Figure 5 provides further insight into how TCMR and LNLN behave under different missing conditions. In the top-right cell (*low missing rate, key semantic cue missing*), the critical sentiment cue "dull" in the phrase "the action was just so dull" is removed. Because LNLN determines its reconstruction strength solely based on the missing rate, the model assigns a relatively small weight to auxiliary modalities, resulting in semantically inconsistent sentiment prediction. In contrast, TCMR recognizes that the key semantic cue is missing and accordingly increases the contribution from other modalities during reconstruction, successfully recovering the negative polarity. The bottom-left cell (*high missing rate, key semantic cue present*) shows the opposite situation: although much of the text is missing, the key cue remains intact. Here, LNLN over-relies on cross-modal information, introducing noise from irrelevant features and ultimately producing an incorrect prediction. TCMR, on the other hand, detects that the key semantic cue is still present and focuses more on the textual modality rather than unnecessary reconstruction, leading to a consistent and correct sentiment inference.

Ablation Study To verify the contribution of each component in the TCMR framework, we conduct an ablation study on both the MOSI and MOSEI datasets. As shown in Table 3, we conduct experiments by removing the AOS, IPFG, and TCE modules one at a time.

In the w/o AOS setting, we consider three variants: `ConfNet-first`, `End2End`, and `ConfNet-later`. Specifically, `ConfNet-first` initially trains the parameters associated with Θ^{conf} to follow the pseudo confidence label c, and then freezes the shared parameters θ^{bert} while optimizing the remaining parameters Θ^{other}. In contrast, `ConfNet-later` begins by optimizing Θ^{other} using the target confidence label c, after which it freezes θ^{bert} and trains only Θ^{conf}. Finally, `End2End` jointly trains all TCMR parameters without any staged optimization.

As shown in Table 3, the TCMR (Full) model achieves the best performance on both MOSI and MOSEI. Among the variants, `ConfNet-later` is the most competitive alternative on MOSI, yet it becomes the worst-performing variant on MOSEI. We attribute this discrepancy to the distinct characteristics of the datasets and the role of the shared BERT encoder during training. From the perspective of dataset difficulty, MOSEI demands broader generalization due to its substantially more diverse topics and speakers than MOSI. Such diversity requires sufficient parameter optimization to learn a much wider range of patterns. However, in the `ConfNet-` variants, the shared BERT encoder is frozen while only Θ^{conf} is optimized, resulting in performance degradation in MOSEI. In contrast, the TCMR (Full) model allows the shared BERT encoder to be optimized in

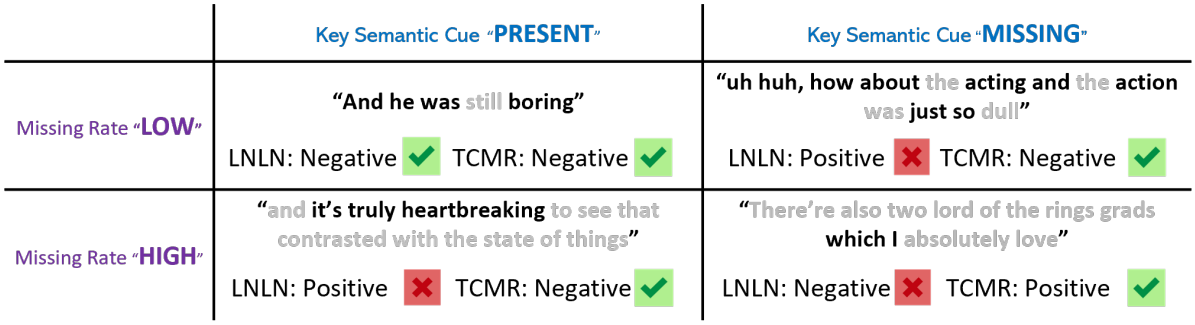


Figure 5. Qualitative comparison between the LNLN and TCMR models under varying missing-text conditions. Each cell presents an example utterance where key semantic cues are either present or missing.

Table 3. A comprehensive ablation study of the proposed TCMR framework on the MOSI and MOSEI datasets, evaluating the contribution of each module (TCE, IPFG, AOS). Note: A smaller MAE indicates better performance.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Non0 Acc / F1	Has0 Acc / F1	MAE	Corr	Acc-7	Acc-5	Non0 Acc / F1	Has0 Acc / F1	MAE	Corr
w/o AOS												
ConfNet-first	27.67	30.65	66.88 / 66.84	66.45 / 66.29	1.1671	43.06	42.13	42.13	64.41 / 58.61	68.67 / 64.42	0.8119	24.07
End2End	29.68	32.07	68.15 / 68.04	67.58 / 67.36	1.1849	49.15	46.82	47.75	77.97 / **77.43**	76.27 / 76.44	0.6619	**58.71**
ConfNet-later	32.24	36.17	71.16 / 71.22	70.41 / 70.37	1.0944	50.91	35.71	35.92	60.52 / 60.11	60.74 / 61.69	0.9262	17.92
w/o IPFG	31.31	34.81	69.02 / 68.91	68.65 / 68.42	1.1674	45.89	46.05	47.06	75.92 / 75.80	72.71 / 73.45	0.6816	57.22
w/o TCE	32.06	**36.80**	70.54 / 70.22	69.98 / 69.45	1.1378	50.27	45.91	46.71	76.99 / 76.51	75.53 / 75.64	0.6738	58.35
TCMR (Full)	**32.79**	36.72	**72.01 / 71.66**	**70.93 / 70.49**	**1.0720**	**52.28**	**47.27**	**48.16**	**77.99** / 76.99	**77.59 / 77.21**	**0.6614**	58.63

a more consistent manner with respect to the two closely related objectives and benefits from AOS training, which mitigates gradient conflicts. Through this coordinated optimization, the BERT encoder learns representations that remain robust under the diverse conditions present in MOSEI. This interpretation is further supported by the inferior performance of the `End2End`, indicating that naive joint optimization is insufficient to achieve such alignment.

In the w/o IPFG setting, we consistently observed performance degradation across both datasets. This indicates that IPFG finely adjusts the contribution of each auxiliary modality for incomplete text reconstruction, thereby effectively filtering out unnecessary or redundant information. In the w/o TCE experiments, to maximize comparability against LNLN, we train TCMR following LNLN's approach. Specifically, we replace the lable of `ConfNet` with the missing rate-based completeness (i.e., one minus the missing rate of the text modality). These results show that TCMR (Full) achieves consistently superior performance across most evaluation metrics, thereby demonstrating the practical effectiveness of our confidence-based completeness approach.

6. Conclusion

This paper proposes TCMR, a framework for multimodal sentiment analysis under partial missing-modality conditions that restores the semantics of incomplete text using confidence scores. As its core idea, we introduce TCE, a confidence estimation module that quantifies the semantic informativeness of incomplete text and enables its reconstruction through proxy features generated from auxiliary modalities. Experiments on two benchmark MSA datasets demonstrate that TCMR can robustly handle noisy and incomplete inputs, indicating its potential applicability in real-world scenarios. For future work, we plan to extend TCMR to a wider range of missing-modality conditions, including more challenging low-information scenarios in which essential cues are simultaneously missing across all modalities.

References

[1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, 2018. Association for Computational Lin-

guistics. 5

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, Lake Placid, NY, USA, 2016. IEEE. 3

[3] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*. Curran Associates, Inc., 2019. 3

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL-HLT 2019), Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 3

[5] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21)*, pages 6–15, Montréal, QC, Canada, 2021. Association for Computing Machinery. 1, 2

[6] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 9180–9192, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 5

[7] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, pages 1122–1131, Seattle, WA, USA, 2020. Association for Computing Machinery. 2, 5

[8] Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 685–696, Seattle, United States, 2022. Association for Computational Linguistics. 1

[9] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia, 2018. Association for Computational Linguistics. 2

[10] Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. M-sena: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 204–213, Dublin, Ireland, 2022. Association for Computational Linguistics. 5

[11] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference (SciPy 2015)*, pages 18–25, 2015. 3

[12] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 271–278. Association for Computational Linguistics, 2004. 1

[13] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, 2002. 1

[14] Yunyan Su, Hong Li, Yifeng Wang, He Zhang, and Zhen Chen. Hierarchical text-guided refinement network for multimodal sentiment analysis. *Entropy*, 27(8):834, 2025. 1

[15] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, pages 3722–3729, Lisboa, Portugal, 2022. Association for Computing Machinery. 2

[16] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, 2019. Association for Computational Linguistics. 2

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 6000–6010, Long Beach, California, USA, 2017. Curran Associates Inc. 3

[18] Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, pages 1–13, 2022. 1, 2, 5

[19] Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259, 2023. 1, 2, 5

[20] Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 5240–5252, Toronto, Canada, 2023. Association for Computational Linguistics. 1

[21] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020. 4

[22] Wenbo Yu, Hua Xu, Zihan Yuan, and Jun Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10790–10797, 2021. 2, 5

[23] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407, 2021. 1, 2, 5

[24] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 5

[25] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, 2017. Association for Computational Linguistics. 2

[26] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 756–767, Singapore, 2023. Association for Computational Linguistics. 1, 2, 5

[27] Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. 1, 2, 3, 4, 5, 7

[28] Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 22123–22138, Vienna, Austria, 2025. Association for Computational Linguistics. 1, 3, 5

[29] Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22123–22138, Vienna, Austria, 2025. Association for Computational Linguistics. 2

CVPR
#18405

TCMR: Text Confidence-aware Missing Semantic Reconstruction for Incomplete Multimodal Sentiment Analysis

Supplementary Material

A. Detailed Model Parameters

To offer a thorough specification of our architecture design, Table 4 outlines the detailed configurations of all modules within TCMR. The Transformer-based modules are responsible for learning rich unimodal and cross-modal representations by modeling long-range contextual dependencies and token-level interactions. The MLP-based modules focus on task-specific computations such as confidence estimation, gating, and sentiment prediction, yielding the final outputs required by each module from the learned representations.

B. Datasets

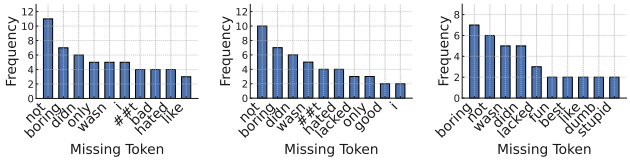
Table 5. Statistics of the MOSI and MOSEI datasets used in our experiments.

Property	MOSI	MOSEI
#Samples	2,199	22,856
Train / Val / Test	1,284 / 229 / 686	16,326 / 1,871 / 4,659
#Speakers	93	1,000+
#Topics	89	250+
Source	YouTube	YouTube
Annotation Type	Utterance-level	Utterance-level
Sentiment Score	–3 to +3	–3 to +3

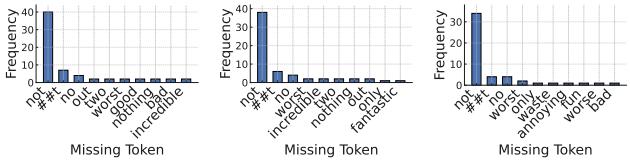
Table 5 provides key statistics of the MOSI and MOSEI datasets used in our experiments. The two datasets exhibit substantial differences in terms of scale and complexity. MOSEI is significantly larger and more diverse in speakers and topics, resulting in greater variability across acoustic, visual, and textual modalities. In contrast, MOSI is comparatively limited in scale and linguistic diversity. Due to this diversity, MOSEI is considered a more challenging benchmark than MOSI [14]. As discussed in Section 5, this characteristic has important implications for model generalization. In addition, the predominance of the neutral class in MOSEI further complicates the model's ability to distinguish subtle positive and negative sentiment cues.

C. Analysis of Pseudo-Confidence Labels

This section provides a more detailed description of the sentiment clue sensitivity analysis introduced in Section 5. Since the pseudo confidence labels serve as the supervisory signal for training the confidence estimator, it is essential to verify that the provided supervision truly reflects the semantic informativeness of the incomplete text.



(a) $c \geq 0.7$ (b) $c \geq 0.8$ (c) $c \geq 0.9$



(d) $c \geq 0.7$ (e) $c \geq 0.8$ (f) $c \geq 0.9$

Figure 6. Token-level sensitivity analysis across different confidence thresholds. Subfigures (a)–(c) present results on the MOSI dataset, while (d)–(f) show the corresponding results on the MOSEI dataset.

Analysis Procedure. To evaluate the sensitivity of pseudo confidence labels to key sentiment tokens, we conduct an analysis following a four-step procedure:

- **Step 1: Correct-sample filtering.** A pre-trained sentiment polarity classifier is applied to the complete textual inputs, and only the correctly classified samples are selected for subsequent steps. Note that a sample is included only when the true-class probability $c \in [0, 1]$ exceeds a predefined threshold.
- **Step 2: Token-wise Augmentation via Masking.** For each correctly classified sample, we generate augmented samples by sliding a window over the sequence and removing exactly one token at each position, replacing it with `[UNK]`. For example, the sentence *"I love this movie"* produces four augmented samples:

  (1) `[UNK]` love this movie.

  (2) I `[UNK]` this movie.

  (3) I love `[UNK]` movie.

  (4) I love this `[UNK]`.

- **Step 3: Re-evaluation of augmented samples.** Then, each augmented sample is passed through the same classifier used in Step 1, and we collect the samples in which masking a specific token causes the predicted sentiment to flip to the opposite polarity (e.g., positive → negative or negative → positive).
- **Step 4: Identification of error-triggering tokens.** Using the polarity-flip samples collected in Step 3, we examine

Table 4. Network configurations of TCMR. Transformer-based modules list the number of layers, sequence lengths, token counts, input dimensions, attention heads, and hidden dimensions for both MOSI and MOSEI (shown as MOSI / MOSEI). MLP-based modules display the input dimensions and hidden-layer widths used in each component.

Module (MOSI / MOSEI)	Type	Notation	#Layers	Seq. len	Token len	Input dim	#Heads	Hidden dim
Transformer-based modules								
BERT Encoder	Transformer	θ^{bert}	12	50 / 50	50	768 / 768	12	768
Text Encoder	Transformer	θ_t^{trans}	2	49 / 49	8	768 / 768	8	128
Audio Encoder	Transformer	θ_a^{trans}	2	375 / 500	8	5 / 74	8	128
Vision Encoder	Transformer	θ_v^{trans}	2	500 / 500	8	20 / 35	8	128
A2T Generator (audio → text)	Transformer	θ_a^{gen}	2	16 / 16	–	128 / 128	8	128
V2T Generator (vision → text)	Transformer	θ_v^{gen}	2	16 / 16	–	128 / 128	8	128
Text Refinement Transformer	Transformer	θ_t^{ref}	2	8 / 8	–	128 / 128	8	128
CrossTransformer (fusion head)	Transformer	θ^{cross}	2	8 / 8	–	128 / 128	8	128
Reconstructor (Audio)	Transformer	θ_a^{recon}	2	8 / 8	–	128 / 128	8	128
Reconstructor (Vision)	Transformer	θ_v^{recon}	2	8 / 8	–	128 / 128	8	128
MLP-based modules								
Confidence Estimator (ConfNet)	MLP	θ^{conf}	6	–	–	768 / 768	–	[768, 768, 1536, 768, 384, 1]
Sentiment Polarity Classifier (Classifier)	MLP	$\theta_{\text{pre}}^{\text{classifier}}$	2	–	–	768 / 768	–	[384, 3]
Gating Network (GateNet)	MLP	θ^{gate}	2	–	–	256 / 256	–	[128, 1]
Text Regressor (Regressor)	MLP	$\theta^{\text{regressor}}$	1	–	–	128 / 128	–	[1]

which token in each flipped sample triggers misclassification and compute how frequently each token leads to a misprediction.

Table 6. Comparison between TCMR and its upper-bound variant TCMR-ub on MOSI and MOSEI.

Dataset	Method	Acc-7	Acc-5	MAE	Corr
MOSI	TCMR	32.79	36.72	1.0720	52.28
	TCMR-ub	37.02	41.24	0.9305	65.03
MOSEI	TCMR	47.27	48.16	0.6614	58.63
	TCMR-ub	57.77	58.72	0.5697	71.78

Analysis Results Figure 6 illustrates which tokens trigger polarity flips when masked under different confidence thresholds. In both datasets, we consistently observed that as the threshold increases key sentiment tokens remain at the top while semantically ambiguous tokens naturally disappear. In MOSI, the token `[I]` does not carry any sentiment meaning, but masking the subject position opens the possibility for negation cues (e.g., don't) to occupy that slot, making the model prone to misprediction. In MOSEI, the token `[two]` is also observed among the flip-triggering words. This is because in the original sentence *"I give the movie two out of five stars"* the token `[two]` serves as an explicit negative cue. When this token is masked, the overall meaning of the sentence *"I give the movie `[UNK]` out of five stars"* becomes ambiguous. Overall, the consistent emergence of key sentiment tokens at higher confidence thresholds suggests that the text classifier is well calibrated with respect to semantic cues.

Table 7. Performance of LNLN and P-RMF when evaluated using checkpoints selected based on test-set performance, following the original training procedure provided in the authors' public code. The values in parentheses indicate the performance difference relative to the corresponding results reported in our main table.

MOSI				
Method	**Acc-7**	**Acc-5**	**MAE**	**Corr**
LNLN	33.45 (+2.09)	37.13 (+2.70)	1.07 (-0.08)	50.35 (+2.22)
P-RMF	31.98 (+3.55)	37.13 (+7.12)	1.08 (-0.05)	50.85 (+2.58)

MOSEI				
Method	**Acc-7**	**Acc-5**	**MAE**	**Corr**
LNLN	46.92 (+0.56)	47.75 (+0.61)	0.66 (-0.01)	58.65 (+0.48)
P-RMF	47.79 (+2.50)	48.78 (+2.60)	0.65 (-0.02)	59.89 (+1.27)

Upper-Bound Analysis To provide a more definitive assessment of the pseudo-labels, we introduce a TCMR variant named TCMR-ub. Specifically, TCMR-ub is trained by directly using the pseudo-labels instead of training the confidence estimator, enabling us to assess the upper-bound performance of our framework. As shown in Table 6, TCMR-ub consistently outperforms TCMR across all metrics. This indicates that the pseudo-labels c provide a meaningful and effective signal for guiding the reconstruction process.

D. Reproducibility Analysis of Baseline Models

To ensure a fair and reproducible comparison with prior work, the performance of LNLN and P-RMF was re-

CVPR
#18405

evaluated using the authors' public implementations rather than relying on the results reported in their papers. During the reproduction experiments, we found that substantial differences emerged between our reproduced results and the original reported results. Through an examination of the released code, we found that both models select their best-performing checkpoints based on the test dataset, which can unintentionally bias the model toward the test dataset. To avoid such concerns, we re-trained all models under a standardized and commonly accepted protocol in machine learning, where model selection is performed using validation-set performance and the test set is used only for final evaluation. Table 7 presents the results obtained by reproducing the two models using their original training procedure, demonstrating that the discrepancies with our main results stem from the test-set-based checkpoint selection rather than from reproduction failures.