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

Anonymous ACL submission

Abstract

Recent advances in multimodal sentiment analysis have increasingly adopted text-centric fusion approaches to effectively leverage the rich sentiment information inherent in the textual modality. Although these approaches have achieved substantial performance improvements, they often suffer from performance degradation when data contain missing values or noise, particularly when sentiment-related cues in the text are missing. To this end, we introduce a novel completeness estimation method that estimates the degree of semantic information loss in incomplete text and leverages it to guide the reconstruction of missing semantics. Furthermore, we propose a training strategy for stable learning in a multi-task setting that jointly optimizes sentiment prediction and completeness estimation. Extensive experiments on three benchmark datasets demonstrate that the proposed approach outperforms existing reconstruction-based MSA models under missing modality conditions.

1 Introduction

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 (Hazarika et al., 2022; Wei et al., 2023), recent MSA studies have increasingly explored text-centric approaches (Han et al., 2021; Wang et al., 2022, 2023; Zhang et al., 2023) 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 when modalities are partially missing, known as the *missing modality problem*.

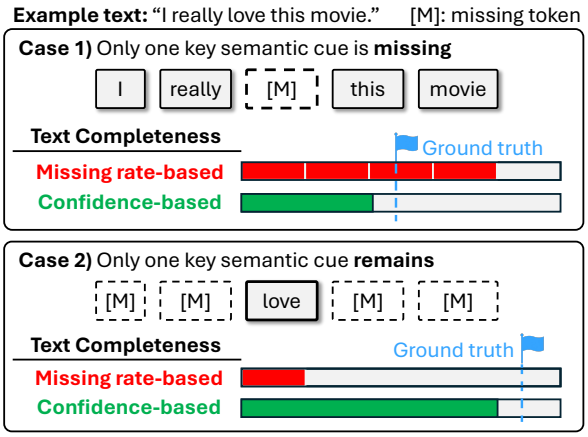


Figure 1: Comparison between missing rate–based and confidence–based completeness. Note that the former is computed as the complement of the missing rate, while the latter is estimated by our proposed method.

To address this issue, several reconstruction-based approaches have been proposed (Yuan et al., 2021; Zhang et al., 2024b; Zhu et al., 2025a). Among them, Zhang et al. (Zhang et al., 2024b) propose LNLN, which estimates text completeness based on the missing rate and uses it as a reconstruction weight to recover missing textual semantics from auxiliary modalities. However, relying solely on the missing rate may fail to accurately reflect semantic informativeness. As illustrated in Figure 1, two contrasting cases highlight this limitation. In Case 1, a missing key sentiment cue (e.g., 'love') leads to substantial semantic loss despite a low missing rate, resulting in an overestimation of completeness. In Case 2, completeness is underestimated even though the sentiment meaning remains intact. This suggests that effective reconstruction requires considering what sentiment information actually remains in incomplete text, rather than relying solely on the missing rate.

Motivated by the above observations, we propose a sentiment-aware completeness estimation approach for more reliable reconstruction of miss-

ing textual semantics. Specifically, we first estimate the completeness of incomplete text through a semantic completeness estimator. To supervise the completeness estimator, we introduce a pseudo-labeling strategy to generate completeness labels. Meanwhile, the Importance-aware Proxy Feature Generator (IPFG) generates proxy features from the auxiliary modalities by adaptively weighting their contributions for text reconstruction. The estimated completeness is then used as a weighting factor to adaptively combine proxy features and observed text features, producing reconstructed textual representations. Lastly, text-guided multimodal fusion is performed using the reconstructed text features as the dominant modality to predict the final sentiment score. In addition, we propose an Alternative Optimization Strategy (AOS) to mitigate gradient conflicts in multi-task learning between the main sentiment prediction task and the completeness estimation task. The contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to introduce a sentiment-aware completeness estimation approach that quantifies the degree of semantic preservation in incomplete text.
- We enhance reconstruction performance through importance-aware proxy feature generation and improve training stability by mitigating gradient conflicts with a task-specific training strategy.
- Through extensive experiments against competitive baselines, we demonstrate that the proposed method consistently outperforms existing reconstruction-based approaches, validating the effectiveness of the proposed completeness label under random missing-modality conditions.

2 Related Work

Recent studies in MSA have explored various multimodal fusion strategies to integrate textual, acoustic, and visual information (Zadeh et al., 2017; Liu et al., 2018; Tsai et al., 2019; Hazarika et al., 2020; Yu et al., 2021; Sun et al., 2022). These approaches aim to capture cross-modal interactions and leverage complementary information across modalities to improve sentiment prediction. Another line of work focuses on text-centric approaches (Han et al., 2021; Wang et al., 2022, 2023; Zhang et al., 2023), which treat text as the dominant modality and utilize audio and visual signals as auxiliary sources to enhance textual representations. However, most existing methods assume that all modalities are fully available during inference, which limits their robustness in real-world scenarios. This issue becomes particularly critical for text-centric approaches, as the loss of important semantic information in text directly undermines the effectiveness of text-centric fusion.

To mitigate the above issue, recent studies have explored reconstruction-based approaches that aim to restore semantic information from missing modalities (Yuan et al., 2021; Zhang et al., 2024b; Zhu et al., 2025b; Wang and Wang, 2025; Yang et al., 2025a,b; Lin and Hu, 2023; Shi et al., 2025; Li et al., 2025; Lin et al., 2025). In particular, LNLN (Zhang et al., 2024b) 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. P-RMF (Zhu et al., 2025b) 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. Additionally, TF-Mamba (Li et al., 2025) proposes a text-enhanced Mamba-based framework that reconstructs missing textual semantics by aligning and enhancing auxiliary modalities through a text-aware modality enhancement module while modeling multimodal dependencies via efficient Mamba-based fusion. While these methods improve robustness to missing modalities, they do not explicitly account for the semantic information loss in incomplete text.

3 Proposed Method

3.1 Problem Definition

The MSA task aims to predict a sentiment score from multiple modalities, including text (t), audio (a), and vision (v) derived from the same utterance.

3.2 Framework Overview

In this work, we introduce a **T**ext **C**ompleteness-based **M**issing **R**econstruction (**TCMR**) framework that estimates textual completeness from incomplete text and reconstructs the missing semantic information using auxiliary modalities. The overall workflow of TCMR is illustrated in Figure 2.

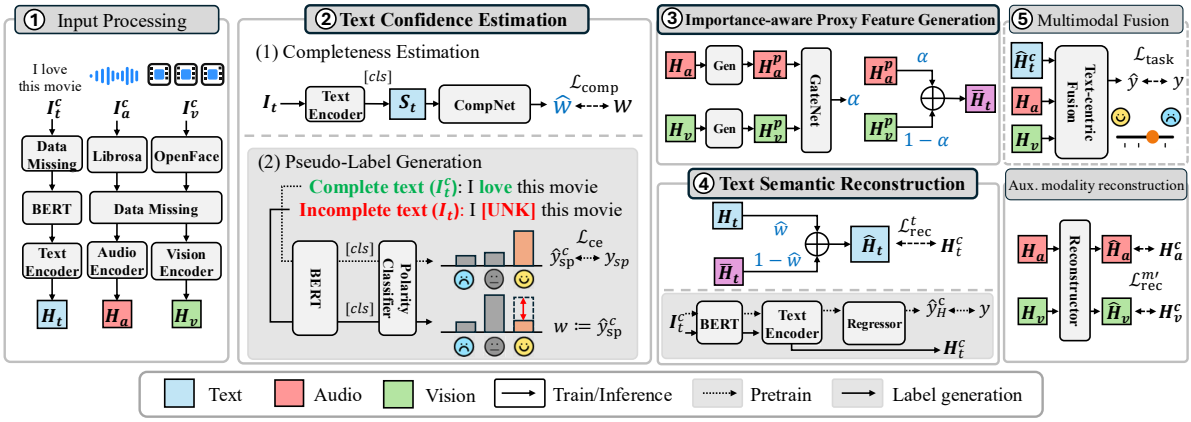


Figure 2: Overview of the proposed framework. Note: Models in the gray boxes are separate and not part of TCMR.

First, each modality input is randomly masked and then embedded into feature representations via modality-specific encoders (①). Next, a completeness estimator predicts how much semantic information remains in incomplete text and reconstructs missing semantics from auxiliary modalities using the predicted value as a weight (②–④). Lastly, a multimodal fusion module performs text-centric fusion to produce the final sentiment score (⑤). Note that the multimodal fusion module is adopted from LNLN, as designing a new fusion module is not the main objective of this work (see Appendix G).

3.3 Input Processing

Raw multimodal inputs are processed into high-level embeddings (Figure 2 ①). First, each raw multimodal input I_m^c for each modality $m \in \{v, a, t\}$, is processed by modality-specific feature extractors: BERT for text, Librosa for audio, and OpenFace for vision (Devlin et al., 2019; McFee et al., 2015; Baltrušaitis et al., 2016).

Following prior works (Zhang et al., 2024b; Zhu et al., 2025a), we adopt a random missing scenario to generate incomplete features $X_m \in \mathbb{R}^{T_m \times d_m}$, where T_m and d_m denote the sequence length and feature dimension. For text, a proportion of tokens in I_t^c are randomly replaced with `[UNK]` according to a missing rate sampled from $[0, 1)$. For audio and vision features, a proportion of the sequence is replaced with zero vectors. Subsequently, the incomplete features are processed by modality-specific Transformer encoders (Vaswani et al., 2017) to obtain the high-level embeddings $H_m \in \mathbb{R}^{T \times d}$:

$$H_m = \text{Transformer}(X_m; \theta_m^{trans}) \tag{1}$$

3.4 Missing Semantic Reconstruction

In this section, we elaborates on the completeness–guided semantic reconstruction pipeline of TCMR (Figure 2 ②–④), which is the core contribution of our work.

Text Completeness Estimation To estimate the completeness of incomplete text, we design a semantic completeness estimator `CompNet`, which consists of fully connected layers followed by a sigmoid activation (Figure 2 ② –1).

Given the BERT `[CLS]` embedding $S_t \in \mathbb{R}^d$, `CompNet` estimates a completeness weight $\hat{w} \in [0, 1]$:

$$\hat{w} = \text{CompNet}(S_t; \theta^{\text{comp}}), \tag{2}$$

The model is trained to minimize mean squared error between the predicted score $\hat{w}$ and the completeness label w:

$$\mathcal{L}_{\text{comp}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{w}^{(i)} - w^{(i)} \right\|^2, \tag{3}$$

where i denotes the data sample index.

Pseudo-Label Generation In our study, we define textual completeness as the extent to which sentiment-related semantic information is preserved in incomplete text. Following this definition, we introduce a pseudo-labeling strategy to generate completeness labels w for training `CompNet` (Figure 2 ② –2).

To quantify the completeness, we extend the concept of True Class Probability (TCP) (Corbière et al., 2019). TCP is commonly used as a confidence measure for classification models, defined as the probability assigned to the ground-truth class.

Our hypothesis is that if a classifier trained on complete text assigns a high TCP to a complete text, the TCP for the corresponding incomplete text will remain high as long as sentiment-relevant tokens are preserved. Conversely, if key sentiment tokens are missing, the TCP may decrease (gray box in Figure 2 ②–2).

Based on this insight, we introduce a sentiment polarity classifier that takes only the text modality as input. First, the classifier is trained on the complete text I_t^c:

$$\hat{y}_{sp} = \text{Classifier}(S_t; \theta^{\text{cls}}) \tag{4}$$

$$\mathcal{L}_{ce} = -\sum_{c=1}^{C} y_{sp}^{(c)} \log \hat{y}_{sp}^{(c)} \tag{5}$$

where y_{sp} denotes the polarity class obtained by discretizing the continuous sentiment score according to predefined thresholds. After training, we perform inference on incomplete text using the pretrained classifier, and the probability assigned to the ground-truth polarity class y_{sp} is used as the pseudo completeness label $w \in [0, 1]$:

$$w^{(i)} \triangleq p\left(y_{\text{sp}}^{(i)} \mid S_t; \theta_{\text{pre}}^{\text{classifier}}\right), \tag{6}$$

where $\theta_{\text{pre}}^{\text{classifier}}$ are the pretrained classifier parameters. The definition of sentiment polarity classes is detailed in Appendix D.1.

Proxy Feature Generation Prior work (Zhang et al., 2024b) generates proxy features without considering the relative importance of auxiliary modalities for reconstruction. However, the contributions of each modality to the reconstruction can vary under random missing conditions.

Therefore, we design an Importance-aware Proxy Feature Generator (IPFG), which consists of Transformer-based modality-specific generators and an MLP-based gating network with a sigmoid:

$$H_{m'}^{p} = \text{Gen}(H_{m'}; \theta_{m'}^{\text{gen}}), \tag{7}$$

$$\alpha = \text{Gate}([H_v^p, H_a^p]; \theta^{\text{gate}}), \tag{8}$$

$$\bar{H}_t = \alpha \odot H_v^p + (1 - \alpha) \odot H_a^p. \tag{9}$$

where $m' \in \{v, a\}$ denotes the auxiliary modalities, $\odot$ denotes element-wise multiplication, and $\alpha \in [0, 1]$ is a gating weight that adaptively balances the contributions of the auxiliary modalities.

Semantic Reconstruction The reconstructed text representation $\hat{H}_t \in \mathbb{R}^{T \times d}$ is generated by integrating the incomplete text feature H_t with the generated proxy feature $\bar{H}_t$, guided by the predicted completeness weight $\hat{w}$ from `CompNet`:

$$\hat{H}_t = \hat{w} H_t + (1 - \hat{w}) \bar{H}_t. \tag{10}$$

To encourage the reconstructed features to be similar to the corresponding complete features, we minimize the mean squared error loss $\mathcal{L}_{rec}^{t}$. For the auxiliary modalities, following prior work (Zhang et al., 2024b), we adopt a self-reconstruction method that takes $H_{m'}$ as input and predicts the corresponding complete representation $\hat{H}_{m'}$. The reconstruction losses are defined as:

$$\mathcal{L}_{\text{rec}}^{m} = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{H}_m^{(i)} - H_m^{c,(i)} \right\|^2, \tag{11}$$

where $H_m^{c,(i)}$ denotes the complete feature of modality m obtained from a model trained with the corresponding complete input (gray box in Figure 2 ④).

3.5 Training Objective

TCMR predicts the final sentiment score $\hat{y}$ through the text-centric multimodal fusion module. The model is trained using the following loss:

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

The overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{comp}} + \beta \mathcal{L}_{\text{rec}}^{t} + \gamma \mathcal{L}_{\text{rec}}^{m'} + \sigma \mathcal{L}_{\text{task}}. \tag{13}$$

where α, β, γ, and σ are hyperparameters.

3.6 Alternative Optimization Strategy

We found that naively minimizing $\mathcal{L}_{\text{total}}$ in an end-to-end manner often leads to unstable training, where $\mathcal{L}_{\text{comp}}$ is effectively ignored and `CompNet` remains stuck at a poor local minimum. We attribute this behavior to gradient conflict (Yu et al., 2020a). This phenomenon commonly arises in multi-task learning, where the gradients of a dominant task suppress the optimization of relatively weaker tasks. In our case, $\mathcal{L}_{\text{task}}$ can also be minimized through shortcut solutions that bypass learning completeness estimation, causing the optimization of $\mathcal{L}_{\text{comp}}$ to be dominated during training.

To address this issue, we introduce a simple yet effective training strategy called Alternative Optimization Strategy (AOS), as illustrated in Algorithm 1. We first partition the total objective $\mathcal{L}_{\text{total}}$ into two groups according to their respective roles:

$$\mathcal{L}_{\text{comp}}, \quad \mathcal{L}_{\text{other}} = \beta\, \mathcal{L}_{\text{rec}}^{t} + \gamma\, \mathcal{L}_{\text{rec}}^{m'} + \sigma\, \mathcal{L}_{\text{task}}. \tag{14}$$

Algorithm 1 Alternative Optimization Strategy

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

for $e = 1$ to E **do**
  Phase 1: optimize Θ^{comp} to minimize $\mathcal{L}_{\text{comp}}$
  for mini-batch $\mathcal{B} \subset \mathcal{D}$, $|\mathcal{B}| = B$ **do**
    $\mathcal{L}_{\text{comp}} \leftarrow \text{LossComp}(\mathcal{B};\ \Theta^{\text{comp}})$
    $\mathcal{L}_{\text{comp}} \leftarrow \alpha\, \mathcal{L}_{\text{conf}}$
    $\text{Opt}_{\text{comp}}.\text{step}\big(\nabla_{\Theta^{\text{comp}}} \mathcal{L}_{\text{comp}}\big)$
  end for
  Phase 2: optimize Θ^{other} to minimize $\mathcal{L}_{\text{other}}$
  for mini-batch $\mathcal{B} \subset \mathcal{D}$, $|\mathcal{B}| = B$ **do**
    $\mathcal{L}_{\text{task}} \leftarrow \text{LossTask}(\mathcal{B};\ \Theta^{\text{other}})$
    $\mathcal{L}^{t}_{\text{rec}} \leftarrow \text{LossRecText}(\mathcal{B};\ \Theta^{\text{other}})$
    $\mathcal{L}^{m'}_{\text{rec}} \leftarrow \text{LossRecAux}(\mathcal{B};\ \Theta^{\text{other}})$
    $\mathcal{L}_{\text{other}} \leftarrow \beta\, \mathcal{L}^{t}_{\text{rec}} + \gamma\, \mathcal{L}^{m'}_{\text{rec}} + \sigma\, \mathcal{L}_{\text{task}}$
    $\text{Opt}_{\text{other}}.\text{step}(\nabla_{\Theta^{\text{other}}} \mathcal{L}_{\text{other}})$
  end for
end for

Within each training epoch, AOS perform a two-phase optimization. In Phase 1, we focus on completeness learning by minimizing $\mathcal{L}_{\text{comp}}$, while fixing the remaining parameters and optimizing only Θ^{comp}. In Phase 2, we fix the learned θ^{comp} and optimize Θ^{other} by minimizing $\mathcal{L}_{\text{other}}$. At this stage, the model focuses on reconstruction and sentiment prediction based on the completeness estimation learned in Phase 1. By alternately optimizing the two objectives, AOS stabilizes the training of `CompNet` and enables the remaining modules to be effectively trained according to their respective objectives. Further discussion of AOS is provided in Appendix C.

4 Experiments

4.1 Experimental Setup

Dataset and Evaluation Metrics We conduct experiments on three MSA benchmark datasets: MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and SIMS (Yu et al., 2020b). In MOSI and MOSEI, sentiment scores are annotated with continuous values ranging from -3 (strongly negative) to $+3$ (strongly positive), while SIMS provides sentiment labels on a scale from -1 (negative) to $+1$ (positive). Model performance is evaluated using Acc-2, Acc-5, Acc-7, F1, MAE, and Corr on MOSI and MOSEI, and Acc-2, Acc-3, Acc-5, F1, MAE, and Corr on SIMS. Dataset statistics and metric definitions are provided in Appendix A.2.

Training and Evaluation Settings Following previous studies (Zhang et al., 2024b; Zhu et al., 2025a), we adopt a *partial random missing* scenario. During training, each modality is randomly erased according to a missing rate r sampled from a uniform distribution over the range $[0, 1.0)$. For each of the three seeds, the best model is selected based on its performance at a missing rate of $r = 0.5$ on the validation set. For testing, we conduct experiments across missing rates ranging from 0 to 0.9 with an increment of 0.1, resulting in ten evaluation settings for each seed.

4.2 Main Experimental Results

Baseline Models We select baseline models with publicly available implementations to ensure reproducibility. For a fair comparison, we reproduce all baseline models using their official implementations. For MISA, Self-MM, MMIM, CENet, TETFN, TFR-Net, and ALMT, we use the MMSA implementations (Mao et al., 2022). For LNLN[1], P-RMF[2] , and TF-Mamba[3], we use their official repositories. In addition, we implement three variants of TCMR with different completeness labeling strategies to analyze their effectiveness:

- **TCMR-TPSC**: estimates completeness using *Target Probability-based Semantic Completeness* (TPSC), as defined in Section 3.4. Note that unless otherwise specified, TCMR refers to TCMR-TPSC for simplicity.
- **TCMR-MRSC**: estimates completeness using *Missing Rate-based Semantic Completeness* (MRSC), which measures completeness according to the missing rate of text.
- **TCMR-LDSC**: To better analyze the effectiveness of TPSC, we design *Lexicon-Derived Semantic Completeness* (LDSC). LDSC generates completeness labels based on sentiment scores of individual tokens obtained from SentiWordNet (Baccianella et al., 2010). Specifically, we first compute the sentiment strength of the complete text by accumulating the sentiment scores of all tokens. Then, the same

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

Table 1: Comparison of the overall performance on MOSI and MOSEI datasets.

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	28.91	31.24	69.91 / 69.57	69.28 / 68.84	1.1302	50.07	47.01	47.82	**78.07 / 77.76**	77.18 / 76.91	0.6667	58.45
TF-Mamba	28.85	31.46	71.05 / 71.11	70.73 / 70.69	1.1228	51.97	45.11	46.09	76.11 / 76.17	72.54 / 73.49	0.6891	57.51
TCMR-MRSC	30.39	33.36	70.84 / 70.95	69.81 / 69.82	1.1061	50.86	46.61	47.46	77.65 / 76.74	76.91 / 76.64	0.6673	58.46
TCMR-LDSC	31.63	35.38	70.61 / 70.55	69.83 / 69.58	1.1322	49.07	47.13	47.87	77.52 / 76.97	76.16 / 76.29	0.6641	58.51
TCMR-TPSC	**32.80**	**36.73**	**72.01 / 71.67**	**70.93 / 70.49**	**1.0721**	**52.28**	47.27	48.16	77.99 / 76.99	**77.59 / 77.21**	**0.6614**	**58.63**

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
MISA	32.97	56.05	71.25	66.95	0.5731	0.319
Self-MM	33.57	55.82	69.02	68.38	0.5243	0.363
MMIM	30.33	55.04	69.46	66.74	0.5629	0.301
CENet	21.61	53.48	69.23	56.97	0.6304	0.014
TETFN	**34.91**	56.01	70.85	69.17	0.5439	0.342
TFR-Net	27.21	54.26	69.37	56.82	0.6258	0.091
ALMT	30.99	53.23	70.65	67.34	0.5445	0.311
LNLN	32.17	56.77	71.51	69.21	0.5423	0.345
P-RMF	32.72	57.30	70.58	**70.36**	0.5432	0.361
TF-Mamba	33.05	55.18	69.88	69.24	0.5235	0.359
TCMR-MRSC	33.27	55.46	70.03	68.76	0.5446	0.341
TCMR-TPSC	30.99	**57.31**	**72.88**	68.87	**0.5222**	**0.374**

Table 2: Comparison of the overall performance on the SIMS dataset. Note: TCMR-LDSC is excluded since it is based on an English sentiment lexicon.

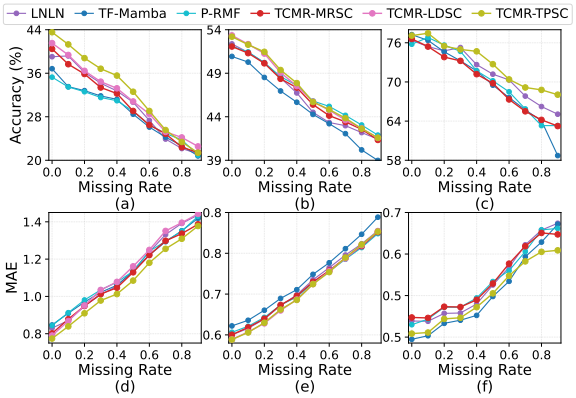


Figure 3: Performance curves under various missing rates. (a)–(c): Accuracy on MOSI (Acc-7), MOSEI (Has0 Acc), and SIMS (Acc-2). (d)–(f): MAE on MOSI, MOSEI, and SIMS.

procedure is applied to the incomplete text. As a result, the completeness label is defined as the ratio between the two strengths. Further implementation details are provided in Appendix B.2.

Overall Performance Table 1–2 present the evaluation results of various models on three MSA benchmark datasets. On the MOSI dataset, TCMR-TPSC achieves state-of-the-art performance across all evaluation metrics. In particular, TCMR-TPSC improves Acc-5 by 6.68% and reduces MAE by 6.37% compared to LNLN. This result verifies that our approach enables more accurate reconstruction, leading to more reliable sentiment prediction. Among the 'TCMR-' variants, TCMR-MRSC shows the worst performance. These results highlight the importance of sentiment-aware completeness estimation. Although TCMR-LDSC achieves competitive performance, TCMR-TPSC consistently performs better across all datasets. This is because LDSC estimates completeness based on sentiment scores assigned to individual tokens, which does not consider contextual semantics and can lead to inaccurate completeness labels. A more detailed analysis is provided in Appendix 7.

On the MOSEI and SIMS datasets, TCMR-TPSC consistently shows strong performance across most evaluation metrics. In particular, on MOSEI, it achieves the best results on Has0 Acc / F1, MAE, and Corr, while remaining competitive on Acc-7, Acc-5, and Non0 Acc / F1. On SIMS, it achieves the best results on Acc-3, Acc-2, MAE, and Corr. Furthermore, to provide a comprehensive comparison with reconstruction-based approaches, we visualize performance at each missing rate in Figure 3. As shown in the figure, TCMR-TPSC consistently outperforms baseline methods, even as the missing rate increases. These results support our motivation that accurately estimating semantic information loss in incomplete text enables effective

Table 3: A comprehensive ablation study of the proposed TCMR framework on the MOSI and MOSEI datasets.

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												
CompNet-first	27.67	30.65	66.88 / 66.84	66.45 / 66.29	1.1671	43.06	43.33	43.48	65.37 / 64.09	69.14 / 65.16	0.7827	32.61
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**
CompNet-later	32.24	36.17	71.16 / 71.22	70.41 / 70.37	1.0944	50.91	43.11	43.86	73.75 / 71.83	74.58 / 73.47	0.7547	46.85
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
TCMR (Full)	**32.80**	**36.73**	**72.01 / 71.67**	**70.93 / 70.49**	**1.0720**	**52.28**	**47.27**	**48.16**	**77.99** / 76.99	**77.59 / 77.21**	**0.6614**	58.63

Table 4: A comprehensive ablation study of the proposed TCMR framework on the SIMS dataset.

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
w/o AOS						
CompNet-first	22.92	53.79	69.82	60.81	0.5764	0.171
End2End	29.26	53.92	68.57	67.71	0.5374	0.321
CompNet-later	29.71	55.49	70.04	68.48	0.5416	0.339
w/o IPFG	30.15	55.57	70.09	66.93	0.5451	0.336
TCMR (Full)	**31.15**	**57.45**	**72.81**	**68.88**	**0.5212**	**0.379**

reconstruction of missing semantic information, thereby leading to improved overall performance.

5 Ablation Study

To verify the contribution of each component, we conduct an ablation study on the MOSI, MOSEI, and SIMS datasets. In the w/o AOS setting, we consider three variants. In `CompNet-first`, we first optimize Θ^{comp}, and then optimize the remaining modules while freezing both the shared BERT encoder and the trained `CompNet`. In contrast, `CompNet-later` first optimizes Θ^{other} using the completeness label w, and then optimizes θ^{comp} while freezing the shared BERT encoder. `End2End` jointly optimizes all model parameters without any staged optimization. In the w/o IPFG setting, we replace IPFG with a simpler proxy feature generator following prior work (Zhang et al., 2024b). Specifically, instead of generating modality-specific proxy features and combining them through a gating mechanism, the audio and visual features are directly used as inputs to a single network that produces the proxy feature.

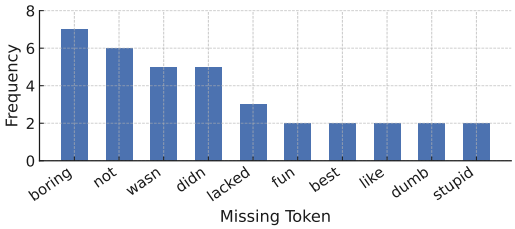
As shown in Table 3, `CompNet-first` consistently achieves the lowest performance across all datasets. One possible reason is that the shared BERT encoder becomes fixed after the completeness learning, which may prevent the model from sufficiently learning text representations for the sentiment prediction task. In contrast, `CompNet-later` achieves relatively competitive performance with TCMR (Full) on MOSI, while it performs relatively worse on MOSEI. We attribute this discrepancy to the distinct characteristics of the datasets and the role of the shared BERT encoder during training. Compared to MOSI, MOSEI contains substantially more samples and covers a wider range of topics. As a result, learning completeness estimation using a relatively small CompNet after the text representations have been optimized may be insufficient to capture the diverse patterns present in MOSEI. Meanwhile, `End2End` shows sub-optimal performance due to the gradient conflict issue discussed in Section 3.6. These observations highlight the necessity of the proposed AOS training strategy for stable optimization.

Removing IPFG consistently degrades performance across all datasets. This suggests that IPFG adaptively regulates the contribution of each auxiliary modality for incomplete text reconstruction, effectively filtering redundant information while enhancing complementary semantic cues.

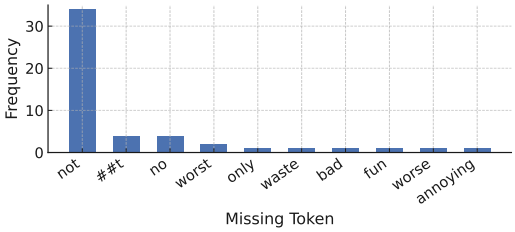
6 In-depth Analysis

Sentiment Token Sensitivity To further examine the effectiveness of the TPSC label, we conduct a sensitivity analysis on sentiment-related tokens. First, we collect samples that are correctly classified by a pretrained sentiment polarity classifier using complete textual inputs. For each sample, we construct a perturbation token set by masking each word token in turn. Then, we count the frequency of tokens whose masking causes a previously correct prediction to become incorrect (detailed procedures are provided in Appendix B).

As shown in Figure 4, misclassifications are concentrated on affective words such as *love*, *hate*, and *boring*. This observation supports our hypothesis that TCP reflects the preservation of sentiment-relevant information, as discussed in Section 3.4. Consequently, `CompNet` provides a more accurate estimate of semantic completeness, enabling more



(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.

reliable reconstruction by maximizing complementary semantic information from auxiliary modalities.

Case Study Figure 5 illustrates how LNLN and TCMR behave under different missing-text conditions. Overall, LNLN either overestimates or underestimates the semantic information remaining in incomplete text, leading to inaccurate reconstruction and ultimately incorrect sentiment prediction. In contrast, TCMR produces correct predictions through more accurate completeness estimation. This is because the estimated completeness appropriately adjusts the balance between auxiliary modalities and the incomplete text, thereby maximizing reconstruction capability. Notably, the bottom-right cell provides an interesting insight. Although LNLN can estimate completeness relatively accurately due to the high missing rate, it still fails to produce the correct sentiment prediction. This suggests that inaccurate completeness estimation during training may limit the learning of the reconstruction process. As a result, accurately estimating the semantic information retained in incomplete text is crucial not only for effective reconstruction but also for stable sentiment prediction.

7 Conclusion

In this work, we are the first to introduce a completeness estimation method for quantifying semantic information loss in incomplete data. Specifically, we propose a TPSC-based pseudo-labeling strategy that reinterprets the target class probability as a measure of completeness. In addition, we introduce LDSC, a lexicon-based completeness metric derived from SentiWordNet, to provide complementary validation of TPSC-based completeness estimation. Through extensive analyses and experiments, we demonstrate that the proposed completeness labels effectively guide semantic reconstruction and improve sentiment prediction performance.

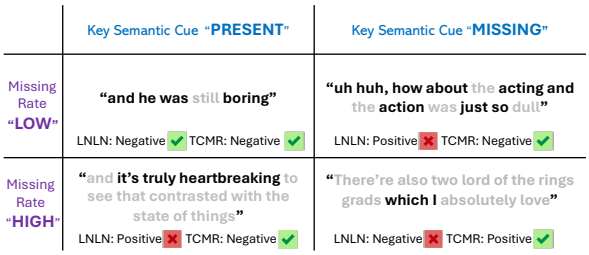


Figure 5: Qualitative comparison between the LNLN and TCMR models under varying missing-text conditions. Each cell presents an example utterance where key semantic cues are either present or missing.

Limitations

Although TCMR consistently outperforms existing reconstruction-based MSA models under random missing conditions, several limitations remain to be addressed in future work. First, the proposed completeness estimation is defined in a text-centric manner. Although TPSC can be extended to other modalities, this study focuses on reconstructing semantic information loss in the textual modality. Extending the TPSC-based completeness estimation to audio and vision modalities may further improve multimodal reconstruction performance and is an important direction for future research. Second, the proposed framework relies on pseudo-labeling, which inherently introduces potential noise in the supervision signal. Although the TPSC-based pseudo labels enable semi-supervised learning by leveraging model predictions on unlabeled data, inaccurate predictions may occasionally result in noisy labels. To mitigate this issue, future work could incorporate confidence-aware filtering strategies. For example, by using only high-confidence TPSC predictions as pseudo labels and integrating complementary labeling strategies such as LDSC for the remaining samples.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2200–2204, Valletta, Malta.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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. Association for Computational Linguistics.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. 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. IEEE.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 794–803.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, volume 32. Curran Associates, Inc.

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 (NAACL-HLT 2019), Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria. 2021. 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. Association for Computing Machinery.

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. 2022. 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. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. 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. Association for Computing Machinery.

Xiang Li, Xianfu Cheng, Dezhuang Miao, Xiaoming Zhang, and Zhoujun Li. 2025. TF-mamba: Text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11252–11267, Suzhou, China. Association for Computational Linguistics.

Ronghao Lin, Qiaolin He, Sijie Mai, Ying Zeng, Aolin Xiong, Li Huang, Yap-Peng Tan, and Haifeng Hu. 2025. Cyclic informative latent space for robust multimodal emotion recognition. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*.

Ronghao Lin and Haifeng Hu. 2023. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. 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. Association for Computational Linguistics.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. 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. Association for Computational Linguistics.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference (SciPy 2015)*, volume 8, pages 18–25.

Piao Shi, Min Hu, Satoshi Nakagawa, Xiangming Zheng, Xuefeng Shi, and Fuji Ren. 2025. Text-guided reconstruction network for sentiment analysis with uncertain missing modalities. *IEEE Transactions on Affective Computing*, 16(3):1825–1838.

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

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. 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. Association for Computing Machinery.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. 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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. 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. Curran Associates Inc.

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

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

Nan Wang and Qi Wang. 2025. Modality-uncertainty-aware knowledge distillation framework for multimodal sentiment analysis. *Complex & Intelligent Systems*, 12(1):1–14.

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. 2023. 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. Association for Computational Linguistics.

Mingzheng Yang, Kai Zhang, Yuyang Ye, Yanghai Zhang, Runlong Yu, and Min Hou. 2025a. Decoupling and reconstructing: A multimodal sentiment analysis framework towards robustness. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 6803–6811.

Mingzheng Yang, Kai Zhang, Yuyang Ye, Yanghai Zhang, Runlong Yu, and Min Hou. 2025b. Decoupling and reconstructing: A multimodal sentiment analysis framework towards robustness. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 6803–6811. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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

Wenbo Yu, Hua Xu, Zihan Yuan, and Jun Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020b. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. 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. Association for Computational Linguistics.

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

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. 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. Association for Computational Linguistics.

Zhi Zhang, Jiayi Shen, Congfeng Cao, Gaole Dai, Shiji Zhou, Qizhe Zhang, Shanghang Zhang, and Ekaterina Shutova. 2024a. Proactive gradient conflict mitigation in multi-task learning: A sparse training perspective. *arXiv preprint arXiv:2411.18615*.

Zhi Zhang, Jiayi Shen, Congfeng Cao, Gaole Dai, Shiji Zhou, Qizhe Zhang, Shanghang Zhang, and Ekaterina Shutova. 2024b. Towards robust multimodal sentiment analysis with incomplete data. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025a. 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. Association for Computational Linguistics.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025b. 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. Association for Computational Linguistics.

A Implementation Details

A.1 Model Configuration

The proposed TCMR model was implemented in PyTorch (v2.2.1) with Python 3.11.7 and trained on a workstation with an NVIDIA GeForce RTX 4090 GPU and an AMD Ryzen Threadripper PRO 5975WX CPU. To offer a thorough specification of our architecture design, Table 5 outlines the detailed configurations of all modules within TCMR.

A.2 Dataset Statistics

Table 6: Statistics of the MOSI, MOSEI, and SIMS datasets used in our experiments.

	MOSI	MOSEI	SIMS
#Samples	2,199	22,856	2,281
Train / Val / Test	1,284 / 229 / 686	16,326 / 1,871 / 4,659	1,368 / 456 / 457
#Speakers	93	1,000+	–
#Topics	89	250+	–
Source	YouTube	YouTube	Movies / TV
Annotation Type	Utterance-level	Utterance-level	Utterance-level
Sentiment Score	–3 to +3	–3 to +3	–1 to +1

Table 6 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 (Su et al., 2025). As discussed in Section 6, this characteristic has important implications for model generalization.

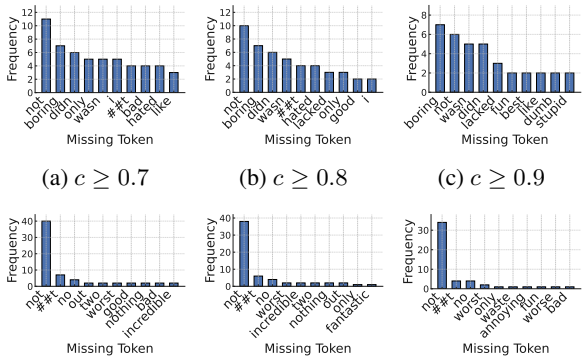
A.3 Evaluation Metrics

Following prior work (Zhang et al., 2024b), we adopt the same evaluation metrics. For MOSI and MOSEI, we report both classification metrics (Acc-7, Acc-5, Non0 Acc / F1, Has0 Acc / F1) and regression metrics (MAE, Corr). The classification metrics are computed by discretizing the continuous sentiment scores into predefined sentiment intervals. For SIMS, we report Acc-2, Acc-3, Acc-5, F1, MAE, and Corr. Non0 Acc denotes negative / positive (excluding zero), whereas Has0 Acc denotes negative / non-negative (including zero).

B Analysis of Completeness Labels

B.1 Sentiment Token Sensitivity of TPSC

This subsection provides a more detailed description of the sentiment token sensitivity analysis introduced in Section 6.



(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 completeness 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

Table 5: 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								
Completeness Estimator (CompNet)	MLP	θ^{comp}	6	–	–	768 / 768	–	[768, 768, 1536, 768, 384, 1]
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]

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

Analysis Results Figure 6 illustrates which tokens trigger polarity flips when masked under different 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.

B.2 Lexicon-derived Completeness Label

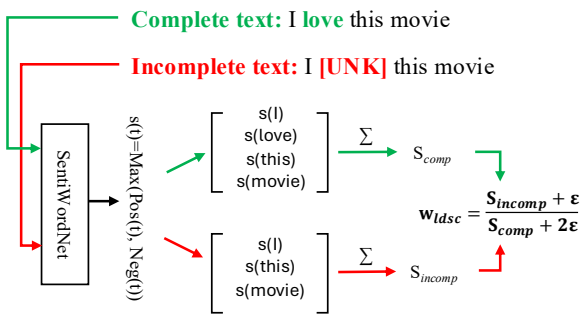


Figure 7: Illustration of LDSC computation

To further validate the effectiveness of the proposed TPSC, we introduce *Lexicon-Derived Semantic Completeness* (LDSC) as a comparison baseline. LDSC measures completeness by explicitly quantifying how much sentiment-related lexical information remains in the incomplete text. Specifically, we adopt SentiWordNet (Baccianella et al., 2010), a widely used lexical resource for sentiment analysis. SentiWordNet provides positive and negative sentiment scores for each word.

Figure 7 illustrates the overall process of generating the LDSC label. First, we define the sentiment intensity score of a token t as:

$$s(t) = \max\big(\text{pos}(t), \text{neg}(t)\big), \tag{15}$$

where $\text{pos}(t)$ and $\text{neg}(t)$ denote the positive and negative sentiment scores of the token. Using this token-level score, we compute the total sentiment-related content of the complete and incomplete text:

$$S_{\text{comp}} = \sum_{t \in T_{\text{comp}}} s(t), \tag{16}$$

Table 7: Qualitative comparison of completeness labels. Each case shows how MRSC, LDSC, and TPSC assign completeness labels depending on whether sentiment-related tokens are missing. In the TPSC column, the value in parentheses denotes $P(y_{sp} \mid \text{complete})$, while the value above it denotes $P(y_{sp} \mid \text{incomplete})$. ✓ and ✗ indicate correct and incorrect completeness estimation, respectively.

Case	Complete Text	Incomplete Text	MRSC	LDSC	TPSC
1	I was happy to see it	I [UNK] happy to see it	0.8333 ✓	0.9888 ✓	0.9526 (0.9580) ✓
2	The soundtrack is good its really good some really great songs	[UNK] soundtrack [UNK] [UNK] [UNK] really [UNK] [UNK] [UNK] [UNK] [UNK]	0.1818 ✓	0.0044 ✓	0.3077 (0.9665) ✓
3	But it was really really awesome	but it was [UNK] [UNK] [UNK]	0.5000 ✗	0.0112 ✓	0.3179 (0.9680) ✓
4	Kids are gonna love the film	[UNK] [UNK] [UNK] love the [UNK]	0.3333 ✗	0.9808 ✓	0.8638 (0.9571) ✓
5	I will admit I'm a big Johnny Depp fan	I will [UNK] I'm [UNK] big Johnny Depp fan	0.7931 ✓	0.0690 ✗	0.9160 (0.9491) ✓
6	and you're better off saving your money uh and maybe renting it when it comes out on uh video so	and [UNK] better [UNK] saving [UNK] money [UNK] [UNK] maybe renting it when it [UNK] out [UNK] uh [UNK] [UNK]	0.5921 ✓	0.9928 ✗	0.4608 (0.6153) ✓
7	I really dig this movie	I [UNK] [UNK] this [UNK]	0.4000 ✗	0.5000 ✗	0.1006 (0.9477) ✓

$$S_{\text{incomp}} = \sum_{t \in T_{\text{incomp}}} s(t), \tag{17}$$

where T_{comp} and T_{incomp} denote the token sets of the complete and incomplete text, respectively. Finally, LDSC-based completeness is computed as:

$$w_{ldsc} = \frac{S_{\text{incomp}} + \varepsilon}{S_{\text{comp}} + 2\varepsilon}, \tag{18}$$

where ε is a small smoothing constant introduced to avoid numerical instability and degenerate cases in which both S_{comp} and S_{incomp} become zero. Intuitively, when $w_{ldsc} \approx 0$, most key sentiment tokens are missing, whereas $w_{ldsc} \approx 1$ indicates that the sentiment-related information is largely preserved.

B.3 Comparison of Completeness Labels

Table 7 provides a qualitative comparison of different completeness labeling strategies under random missing conditions. The results in Cases 3 and 4 highlight the limitations of MRSC. Since MRSC estimates completeness solely based on the proportion of missing tokens, it cannot capture the actual semantic information loss in the incomplete text. In contrast, LDSC and TPSC more appropriately reflect completeness by considering the semantic importance of sentiment-related tokens.

Cases 5–7 reveal the limitations of LDSC. For example, in Cases 5 and 7, words such as [fan] and [dig] are used as positive sentiment expressions. However, WordNet assigns sentiment scores based on the most frequent sense of a word, which may differ from the sense used in the actual context, resulting in incorrect sentiment score assignments. In Case 6, [better] is assigned a positive sentiment score in WordSentiNet, leading LDSC to predict a high completeness value even though the overall sentence conveys negative sentiment. This result highlights a fundamental limitation of lexicon-based approaches, which cannot adequately capture contextual semantics.

C Comparison with Gradient Conflict Baselines

Table 8: Gradient conflict rate (GCR) analysis measured during training. F50/L50 denote the first and last 50% of training epochs, respectively.

Method	GCR (%) ↓			$\mathcal{L}_{comp}$ ↓
	All	F50	L50	
E2E	51.27	51.16	54.36	0.1634
GradNorm	54.84	56.55	53.16	0.1294
PCGrad	52.33	57.43	48.70	0.0340
AOS	48.47	53.67	43.33	0.0055

In multi-task learning, jointly optimizing multiple losses often leads to gradient conflicts, resulting in biased optimization where certain tasks domi-

nate while others are under-optimized. To address this problem, prior work (Chen et al., 2018; Yu et al., 2020a) has proposed methods that mitigate gradient conflicts by adjusting gradient magnitudes or directions during training. For example, Chen et al. (Chen et al., 2018) propose GradNorm, a method that automatically balances multiple tasks during training by adjusting the magnitudes of gradients derived from each task-specific loss, ensuring that all tasks learn at a similar rate. Yu et al. (Yu et al., 2020a) propose PCGrad, which reduces gradient interference by projecting conflicting gradients onto the normal plane of other task gradients.

To validate the effectiveness of AOS, we conduct comparative experiments using existing methods as baselines. Table 8 presents the gradient conflict rates during training and the corresponding convergence behavior of the completeness estimation loss, following the protocol of prior work (Zhang et al., 2024a). In joint training, $\mathcal{L}_{\text{comp}}$ tends to converge to a local optimum, which leads to an increase in GCR (see E2E in Table 8). While GradNorm and PCGrad mitigate gradient conflict, they are still insufficient to escape local optima due to the fundamental limitation of joint training. In contrast, AOS removes interference from other losses during the `CompNet` training phase, thereby effectively mitigating suboptimal convergence of $\mathcal{L}_{\text{comp}}$.

D Sentiment Polarity Classes

D.1 Definition of Sentiment Polarity Classes

To utilize TCP, we convert continuous sentiment scores into discrete polarity classes. As the score ranges differ across datasets, the class boundaries are defined separately for SIMS and MOSI/MOSEI.

For the SIMS dataset, polarity classes are defined using fixed interval thresholds. Specifically, the 2-class setting divides the scores into $[-1.0, 0.0]$ and $(0.0, 1.0]$. The 3-class setting introduces a neutral region, resulting in $[-1.0, -0.1]$, $(-0.1, 0.1]$, and $(0.1, 1.0]$. The 5-class setting adopts a finer partition given by $[-1.0, -0.7]$, $(-0.7, -0.1]$, $(-0.1, 0.1]$, $(0.1, 0.7]$, and $(0.7, 1.0]$.

For the MOSI and MOSEI datasets, polarity classes are defined by mapping sentiment scores into discrete levels. The 2-class setting separates negative (< 0) and non-negative (≥ 0) samples. The 3-class, 5-class, and 7-class settings correspond to $\{-1, 0, 1\}$, $\{-2, -1, 0, 1, 2\}$, and $\{-3, -2, -1, 0, 1, 2, 3\}$, respectively.

D.2 Optimal Choice of Class Granularity

Table 9: Performance sensitivity to sentiment class granularity (k) in TPSC label generation

(a) MOSI dataset

Model	Acc-7	Acc-5	Non0 Acc/F1	Has0 Acc/F1	MAE	Corr
TPSC-2cls	31.45	36.01	70.01 / 69.91	69.46 / 69.25	1.1073	50.66
TPSC-3cls	**32.80**	**36.73**	**72.01 / 71.67**	**70.93 / 70.49**	**1.0720**	**52.28**
TPSC-5cls	30.96	35.12	69.51 / 69.09	68.93 / 68.42	1.1449	49.62
TPSC-7cls	32.16	35.35	70.85 / 70.76	69.74 / 69.54	1.1040	49.28

(b) MOSEI dataset

Model	Acc-7	Acc-5	Non0 Acc/F1	Has0 Acc/F1	MAE	Corr
TPSC-2cls	46.41	47.25	77.76 / 77.32	76.31 / 76.53	0.6677	57.95
TPSC-3cls	**47.27**	**48.16**	**77.99** / 76.99	**77.59 / 77.21**	**0.6614**	58.63
TPSC-5cls	46.94	47.96	77.79 / **77.56**	75.46 / 75.95	0.6638	**59.28**
TPSC-7cls	46.75	47.63	77.74 / 76.74	77.15 / 76.79	0.6667	57.81

(c) SIMS dataset

Model	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
TPSC-2cls	30.99	**57.31**	**72.88**	68.87	**0.5222**	**0.374**
TPSC-3cls	32.43	56.23	71.72	**68.96**	0.5478	0.321
TPSC-5cls	**33.14**	55.45	69.13	68.06	0.5465	0.329

We conducted comparative experiments with $k \in \{2, 3, 5, 7\}$ to determine the optimal number of sentiment classes for training TPSC. This conversion of continuous sentiment scores into discrete classes is essential for utilizing the model's confidence—specifically, the True Class Probability (TCP)—as a proxy for semantic completeness, which is not directly obtainable from a regression-based framework. By redefining the task as a classification problem, we can quantify the model's certainty regarding the preservation of sentiment information. In this context, the choice of class granularity (k) serves as a critical factor that determines both the sensitivity and stability of the completeness estimation. Detailed experimental results are presented in Table 9.

Experimental results indicate that the optimal number of classes to maximize model robustness is $k = 3$ for MOSI/MOSEI and $k = 2$ for SIMS. For MOSI and MOSEI, the 3-class configuration yielded the best performance as it effectively captures the semantic value of the neutral zone within their wide sentiment ranges. Conversely, for the SIMS dataset, which has a narrower score range, the predictive probability based on binary classification (2-class) proved to be a more reliable and stable indicator of completeness. By adopting these dataset-specific granularities, the TCMR model was able to enhance the precision of semantic reconstruction and achieve more robust sentiment

prediction performance even under incomplete text conditions.

E Training Efficiency Analysis

We compare the training overhead of each model by measuring the number of parameters and the training time per epoch. As shown in Table 10, TCMR achieves the second fastest training speed, following TF-Mamba. This is because the reconstruction modules used in LNLN and P-RMF involve more computationally intensive operations. Specifically, LNLN employs a Transformer-based reconstruction network for semantic reconstruction, while P-RMF adopts a VAE-based generative reconstruction framework. In contrast, since TCMR employs MLP-based CompNet, it maintains efficient training speed even though it has a slightly larger number of parameters. As a result, the additional optimization steps introduced by AOS do not significantly increase the overall training cost.

Table 10: Comparison of model complexity and training time per epoch.

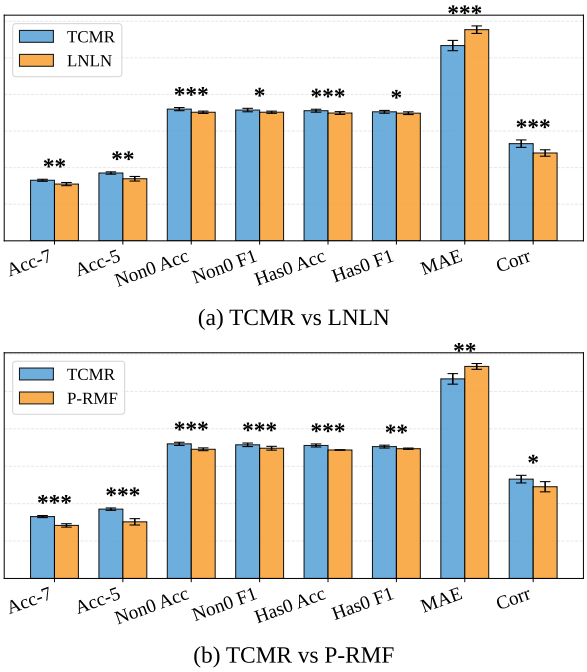
Model	# Params (M)	Time / Epoch (s)
LNLN	115.97	15.40
P-RMF	117.31	18.00
TF-Mamba	111.30	11.12
TCMR	119.23	12.48

F Statistical Analysis

To examine the statistical significance of performance differences among models, we conduct comparative experiments on the MOSI dataset with LNLN, P-RMF, and TCMR. As shown in Figure 8, TCMR achieves statistically significant performance improvements across most evaluation metrics.

G Multimodal Fusion and Sentiment Prediction

The multimodal fusion module follows the design of a prior work (Zhang et al., 2024b), and its key components are briefly summarized here. Starting from the reconstructed text representation H_t^p, a refinement Transformer encoder applies self-attention, producing layerwise refined features $H_t^{+(i)}$ with $i \in \{1, \ldots, l^{\text{ref}}\}$. 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



(a) TCMR vs LNLN

(b) TCMR vs P-RMF

Figure 8: Error bars indicate standard deviations. Asterisks denote statistical significance (* p < 0.1, ** p < 0.05, *** p < 0.01).

feature $H_t^{+(i)}$ serves as the query and attends to H_v and H_a, while the fused representation updates $H_f^{(i)}$ through residual accumulation. In compact form,$H_f^{(i)} = H_f^{(i-1)} + \text{MHA}(H_t^{+(i)}, H_v) + \text{MHA}(H_t^{+(i)}, H_a)$ 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 $\hat{y}$.