Controlling Decision Drift in Multimodal Sentiment Analysis with Missing Modalities

Chenglizhao Chen[1,2], **Yuchen Cao**[1,2], **Xinyu Liu**[1,2*], **Mengke Song**[1,2], **Guisheng Zhang**[1,2] and **Xiaomin Yu**[3]

[1]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China)

[2]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software

[3]The Hong Kong University of Science and Technology (Guangzhou)

{cclz123, cyc021109, liuxy001005, songsook, sdut_guisheng}@163.com, yuxm02@gmail.com

Abstract

Multimodal sentiment analysis relies on textual, acoustic, and visual signals, yet real-world data often suffer from modality missing and quality imbalance. Existing methods generate features for modality missing from available ones, but differences in expression mechanisms and sentiment dynamics across modalities may cause the generated features to deviate from true distributions and mislead prediction. In addition, unreliable modalities may dominate fusion, resulting in representation shift across modality combinations and unstable sentiment representations. To address these challenges, we propose a two-level reference alignment framework. The framework introduces stable references at the feature representation and sentiment decision levels to improve robustness under modality missing. First-level reference alignment leverages complete-modality samples to constrain representations and align different modality combinations into a shared sentiment space. Second-level reference alignment enforces cross-modal consistency at the decision level by suppressing unreliable modalities through prototype retrieval and voting. As a result, the framework maintains stable and reliable sentiment predictions under diverse missing-modality patterns. Experiments on CMU-MOSI and CMU-MOSEI show consistent improvements across various missing-modality settings. Under full-modality input, the proposed method achieves state-of-the-art performance, with ACC of 86.28% and 85.88%, and F1 of 86.24% and 85.86%.

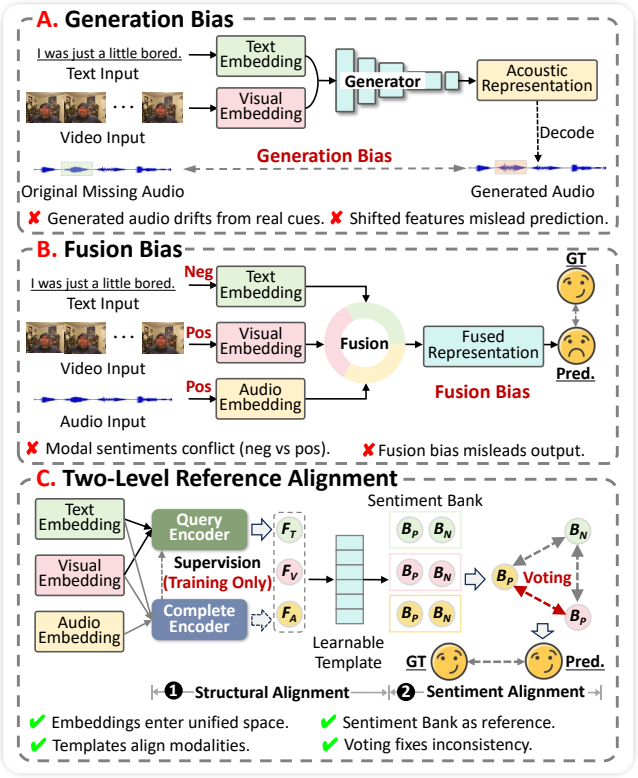

Figure 1: Illustration of generation bias, fusion bias, and Two-Level Reference Alignment. (A) Missing modalities shift generated features away from sentiment cues. (B) Unreliable modalities dominate fusion and mislead prediction. (C) Feature- and decision-level references align representations and reduce both biases.

1 Introduction

The objective of sentiment analysis is to identify human sentiments during communication, which provides a foundation for understanding behaviors and social interactions [Liu, 2022]. Sentiments are expressed through textual content, vocal intonation, and body language, making multimodal sentiment analysis necessary for capturing emotional dynamics more comprehensively [Poria *et al.*, 2017]. Integrating textual, acoustic, and visual information enables richer affective representation and supports broader human-computer interaction applications [Zadeh *et al.*, 2018b].

Specifically, in real-world deployments, multimodal observations are often missing or unreliable due to environmental conditions, occlusions, and system failures [Wu *et al.*, 2024;

*Corresponding author.

Reza *et al.*, 2024]. Such unreliability may arise from background speech, motion blur, illumination variation, occlusion, or sensor failure, which commonly lead to unstable representations and biased fusion decisions. These situations introduce inconsistent cues and make it difficult to obtain accurate sentiment representations [Yu *et al.*, 2021]. Under modality missing scenarios, many methods generate missing-modality features from available modalities to preserve the multimodal prediction structure [Guo *et al.*, 2024]. However, due to differences in expression mechanisms and sentiment dynamics, generated features may deviate from the true feature space and incur generation bias, as illustrated in Figure 1(A) [Dai *et al.*, 2025], further misleading fusion and drifting predictions toward incorrect semantic regions [Liu *et al.*, 2024].

Even when all modalities are available, multimodal systems still face quality imbalance. Modalities affected by noise, occlusion, or abnormal expressions may be treated as reliable signals during fusion [Arevalo *et al.*, 2017; Mao *et al.*, 2023]. Although such corruptions differ at the input level, they commonly reduce modality reliability and introduce inconsistent sentiment evidence. As shown in Figure 1(B), unreliable sources may be over-weighted in the fused representation, pushing predictions away from the correct sentiment target and causing fusion bias [Baltrušaitis *et al.*, 2018].

To address generation bias and fusion bias, we introduce a two-level reference alignment (TLRA) framework that provides controllable cross-modal references at both the feature and sentiment decision levels, as illustrated in Figure 1(C). TLRA uses complete-modality information as stable references to guide feature alignment under modality missing, thereby reducing the deviation of generated or completed features from the true sentiment space. It further models cross-modal consistency at the decision level through reference-driven aggregation, suppressing the influence of unreliable modalities on final prediction. Rather than designing noise-specific correction modules, TLRA treats missing and unreliable modalities as a unified reliability degradation problem and constrains their effects through reference alignment.

The contributions of this work are:

- A TLRA framework that introduces cross-modal references at both feature and sentiment decision levels to address generation bias under modality missing and fusion bias caused by unreliable modalities.
- A reference-constrained feature alignment strategy that leverages complete-modality samples as stable semantic anchors to reduce generation bias and stabilize representations under modality missing.
- A reference-driven decision alignment strategy that models sentiment-level cross-modal consistency with prototype-guided references and reduces the influence of potentially unreliable modalities during fusion.
- Extensive experiments on CMU-MOSI and CMU-MOSEI demonstrate that the proposed method outperforms representative methods under various missing-modality scenarios and achieves state-of-the-art performance under complete modalities. [1]

[1]https://github.com/LiuXY3366/TLRA

2 Related Works

2.1 Sentiment Analysis

Early multimodal sentiment analysis methods commonly assume complete and well-aligned modalities, improving affective modeling through cross-modal attention, feature interaction, and temporal fusion [Liu *et al.*, 2018; Tsai *et al.*, 2019; Yu *et al.*, 2024]. However, real-world modalities can be degraded by noise or missing values, making these methods sensitive to modality stability [Baltrušaitis *et al.*, 2018; Zeng *et al.*, 2022; Li *et al.*, 2024]. Existing studies address this issue using cross-modal generation, text-dominant recovery, or robustness training [Neverova *et al.*, 2015; Pham *et al.*, 2019; Lin and Hu, 2023]. Nevertheless, because modalities differ in expression mechanisms and temporal dynamics, recovered features may not align with the true sentiment structure and can produce unstable or conflicting fused representations [Lin and Hu, 2023; Zeng *et al.*, 2022]. Unlike generation-based recovery, our method introduces stable cross-modal references to constrain sentiment representations under missing or quality-imbalanced modality conditions.

2.2 Multimodal Fusion

In multimodal sentiment analysis, fusion performance is strongly affected by modality quality, expression stability, and signal conditions [Liu, 2022]. Prior studies improve robust fusion by estimating modality confidence, adaptively weighting modality contributions, or imposing consistency objectives to reduce uncertainty-induced drift [Arevalo *et al.*, 2017; Zadeh *et al.*, 2018a; Wang *et al.*, 2019; Xie *et al.*, 2024; Han *et al.*, 2021; Hazarika *et al.*, 2020]. In contrast, this work introduces prototype-based decision references to align modality-wise predictions and enhance cross-modal consistency when modality quality varies in practical scenarios.

3 Method

3.1 Overview

To mitigate both generation bias from modality missing and fusion bias from quality imbalance, we develop the Two-Level Reference Alignment framework (TLRA). TLRA defines cross-modal references at both the representation and decision levels, allowing the model to learn consistent and robust sentiment representations even when only partial or unreliable multimodal inputs are available.

As illustrated in Figure 2, TLRA consists of three key components: modality encoding, representation-level alignment, and decision-level alignment. For textual, acoustic, and visual modalities, TLRA independently encodes available modality signals while constructing sentiment reference structures from complete-modality training samples. The representation-level alignment stage constrains incomplete features into a shared sentiment space to reduce generation-induced representation drift. The decision-level alignment stage further enforces cross-modal consistency by suppressing unreliable modality decisions during fusion, thereby reducing fusion bias and improving prediction stability.

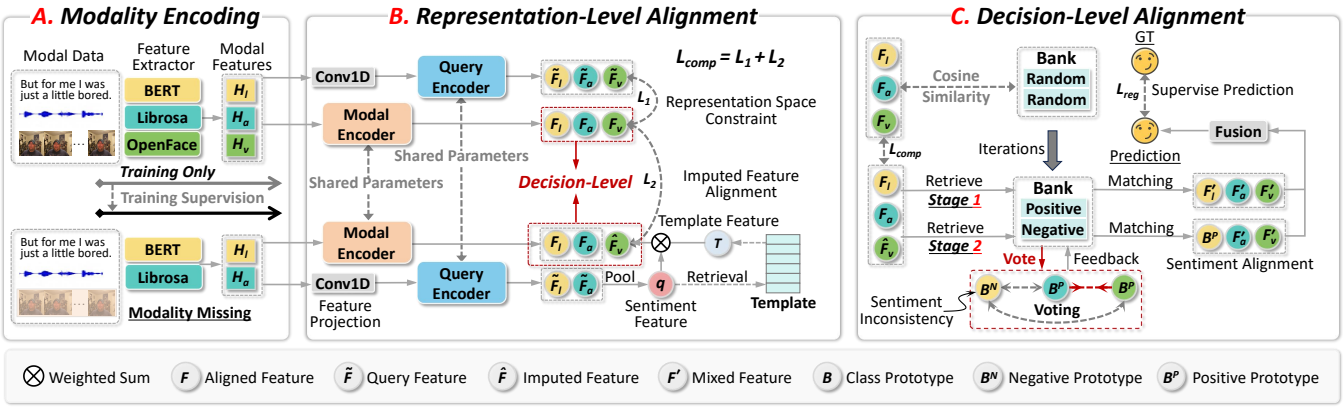

Figure 2: Overall pipeline of TLRA. The framework consists of three stages: (A) Modality Encoding, (B) Representation-Level Alignment, and (C) Decision-Level Alignment. It first extracts modality-specific features, then aligns incomplete representations via reference-driven completion, and finally performs prototype-guided decision alignment to ensure robust sentiment predictions under modality missing.

3.2 Modality Encoding

In the modality encoding stage, we independently model the raw signals of each modality based on the presence of potentially missing modalities, creating a structured representation for each modality to serve as input for future alignment. Figure 2(A) depicts that during the modality encoding stage, the TLRA framework only encodes the available modalities (i.e., does not complete or reconstruct missing modalities) at this stage of the TLRA framework, which allows us to perform the alignment step under true conditions of modality missing.

For the textual, acoustic, and visual modalities, we utilize the BERT pretrained language model, the Librosa acoustic feature extractor, and the OpenFace visual analysis system, respectively, to encode the modalities [Devlin *et al.*, 2019; McFee *et al.*, 2015; Baltrušaitis *et al.*, 2016]. A multimodal input will result in three modality representations: H_l, H_a, and H_v for the text, audio, and visual modalities, respectively.

During training, complete-modality samples are used as supervision references for constructing reference structures. When one or more modalities are missing, TLRA operates only on available modalities without encoding or reconstructing missing inputs. This design allows the model to distinguish missing information from available evidence and avoids introducing unreliable assumptions at the input level.

3.3 Representation-Level Alignment

Imputed features may deviate from the actual sentiment structure and cause incorrect predictions, mainly due to the lack of stable semantic references for anchoring incomplete inputs. TLRA addresses this issue through reference-driven representation-level alignment, which constrains imputed representations within the semantic space defined by complete modalities, as shown in Figure 2(B).

Dual-Path Encoding Strategy. TLRA uses a dual-path encoding strategy to handle both complete-modality and modality-missing samples. The complete-modality path establishes a stable sentiment space from complete inputs, while the modality-missing path serves as a query and completion path for incomplete inputs. By decoupling reference construction from query-based completion, the model can align missing-modality representations within the same sentiment space without assuming that all modalities are available.

Given modality representations H_m from the encoding stage, where $m \in \mathcal{M}$ denotes available modalities, the Modal Encoder maps each modality into a shared sentiment space:

$$F_m = \mathrm{Enc}_M(H_m), \quad m \in \mathcal{M}. \tag{1}$$

The resulting features F_m exhibit stable sentiment structure and support subsequent decision-level alignment.

To support representation completion under modality missing, a Query Encoder is introduced to construct query features for reference retrieval. Each available modality is first projected into a unified representation space via a lightweight 1D convolution, and then processed by the Query Encoder:

$$\tilde{F}_m = \mathrm{Enc}_Q\big(\mathrm{Conv1D}(H_m)\big), \quad m \in \mathcal{M}. \tag{2}$$

All modalities pass through the query path regardless of completeness, producing features that support both completion and alignment. By decoupling reference construction from query-based alignment in a shared semantic space, the dual-path design enables robust alignment without assuming complete inputs. The two-path encoding is implemented with shared lightweight modules to ensure efficiency and stability.

Template-Guided Completion. Under modality missing, unconstrained feature generation can produce biased representations deviating from true sentiment structures. To address this, TLRA introduces a template-guided completion mechanism that provides stable semantic references for missing modalities and constrains completion within the sentiment space defined by complete modalities.

Given query features $\tilde{F}_m$ from the Query Encoding, we apply temporal pooling to obtain a global query vector q_m:

$$q_m = \mathrm{Pool}(\tilde{F}_m). \tag{3}$$

TLRA maintains a learnable template library $\mathcal{T} = \{t_1, t_2, \ldots, t_K\}$ in the shared sentiment space. The similarity between q_m and each template is computed as:

$$s_k = q_m^{\mathrm{T}} t_k. \tag{4}$$

The similarities are normalized via softmax to obtain aggregation weights α_k, yielding the template context as:

$$T_m = \sum_{k=1}^{K} \alpha_k t_k. \tag{5}$$

To adaptively fuse the sample-specific query and the template reference, we compute a consistency weight based on their cosine similarity as:

$$\alpha_c = \cos(q_m, T_m), \tag{6}$$

The completed representation is then obtained as:

$$\hat{F}_m = \alpha_c \, q_m + (1 - \alpha_c) \, T_m. \tag{7}$$

The completed representation $\hat{F}_m$ serves as sentiment-level structural compensation rather than reconstruction of the original modality signal. Therefore, the completion process focuses on recovering stable sentiment structure instead of synthesizing raw modality-specific details, which reduces the risk of introducing modality-specific generation noise. By constraining completed features within the shared sentiment space, template-guided completion effectively suppresses generation bias under modality missing and provides stable inputs for alignment and fusion.

Representation-Level Alignment Loss. The objective of representation-level alignment is to enforce structural consistency among intermediate representations within a unified sentiment space under modality missing. Specifically, both query features for observed modalities and completed features for missing modalities are aligned to the modality representations defined under complete inputs.

Let $\mathcal{M}$ denote the set of available modalities and $\mathcal{M}_{\text{miss}}$ the set of missing modalities. We define $\Omega = \mathcal{M} \cup \mathcal{M}_{\text{miss}}$ as the set of modalities involved in alignment. Let F_m denote the modality feature produced by the modal encoder, $\tilde{F}_m$ the query feature, and $\hat{F}_m$ the completed feature. The unified representation alignment loss is defined as:

$$\mathcal{L}_{\text{align}} = \lambda_1 \sum_{m \in \mathcal{M}} \left\| \tilde{F}_m - F_m \right\|_2^2 + \lambda_2 \sum_{m \in \mathcal{M}_{\text{miss}}} \left\| \hat{F}_m - F_m \right\|_2^2. \tag{8}$$

By aligning query features and completed missing-modality features to complete-modality references, this loss suppresses representation drift caused by missing inputs. As a result, different modality combinations are encouraged to follow a consistent sentiment structure, providing more reliable inputs for decision-level alignment.

3.4 Decision-Level Alignment

Although representation-level alignment reduces missing-induced drift, decision inconsistencies may still occur when noisy modalities produce misleading predictions. These unreliable modality-wise predictions can dominate fusion and degrade the final sentiment decision. As shown in Figure 2(C), TLRA filters and aligns modality-wise predictions to suppress unreliable evidence under varying modality conditions and improve decision robustness.

Prototype Construction. For each sentiment class c, we maintain modality-specific prototypes B_m^c located within the shared sentiment space for the text, audio, and visual modalities. These prototypes serve as decision-level references and are constructed based upon accumulated training features, capturing stable class structural elements across modalities. In practice, the prototype bank is randomly initialized and then updated online during training, so no extra clustering procedure or external prior is required.

Given the aligned feature F_m of the sample with class label c, we determine the cosine similarity with the associated prototype B_m^c, which is utilized to create a mixed prototype:

$$B'^c_m = \alpha_m F_m + (1 - \alpha_m) B_m^c, \tag{9}$$

where α_m is determined by the cosine similarity between F_m and B_m^c, adaptively balancing sample-specific information and class-level structure to improve robustness.

The mixed prototype B'^c_m enables gradual updates of the prototype memory with controlled momentum, allowing prototypes to evolve into stable class-level references while remaining robust to sample variations, as illustrated in Figure 3(A). In the binary sentiment setting, each modality maintains one positive and one negative prototype, keeping the memory compact and interpretable while providing explicit class-level anchors. This class-wise design avoids additional prototype-number search and ensures that decision references are directly aligned with sentiment labels. During training, we adopt momentum-based updates with normalization to stabilize cosine similarity, enabling prototypes to progressively evolve from coarse initial references into reliable and discriminative sentiment anchors in practice.

Through this process, class prototypes act as decision-level anchors that provide consistent references across samples and modalities. As the prototype memory is continuously updated during training, it gradually captures stable class-level sentiment structures and helps resolve cross-modal prediction conflicts under noisy and unreliable modality conditions.

Two-Stage Decision-Level Alignment. Since prototype memory gradually evolves from coarse class representations into reliable references, TLRA adopts a two-stage decision-level alignment strategy that progressively strengthens cross-modal consistency, as shown in Figure 2(C).

Stage1: Soft Prototype Guidance. At early epochs, prototype memory is not yet stable enough to support hard voting, and strong cross-modal constraints may amplify modality noise. Therefore, prototypes provide continuous sentiment references without explicit voting or hard class decisions. This soft guidance preserves useful modality-specific information while gradually pulling unstable features toward smoother and more reliable prototype structures.

Given the aligned modality feature F_m for each available modality $m \in \mathcal{M}$, the model retrieves the corresponding positive and negative prototypes B_m^P and B_m^N from the prototype memory and computes their cosine similarities:

$$s_m^P = \cos(F_m, B_m^P), \qquad s_m^N = \cos(F_m, B_m^N). \tag{10}$$

The prototype with higher similarity is selected as the pri-

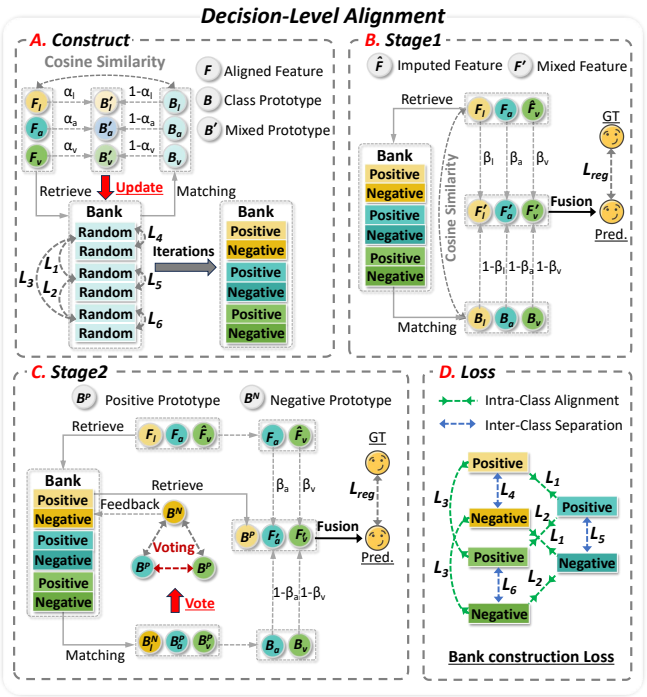

Figure 3: Decision-level alignment in TLRA. It illustrates prototype construction, a two-stage decision alignment strategy, and the training objectives for consistent and robust sentiment predictions.

mary sentiment reference for the current modality:

$$B_m^* = \arg\max_{c\in\{P,N\}} \cos(F_m, B_m^c), \tag{11}$$

and the maximum similarity is used as a consistency score:

$$\beta_m = \max(s_m^P,\ s_m^N). \tag{12}$$

Based on this score, the modality feature is fused with the selected prototype to form a mixed representation F'_m:

$$F'_m = \beta_m F_m + (1-\beta_m) B_m^*. \tag{13}$$

This soft guidance preserves reliable modality information while pulling unstable modalities toward smoother prototype structures. The resulting modality representations are then fused to produce the final sentiment prediction.

Stage2: Prototype Voting and Suppression. As training progresses, the prototype memory gradually becomes compact and discriminative. In the second stage, TLRA introduces an explicit prototype voting mechanism to enforce cross-modal decision consistency, as shown in Figure 2(C). Each modality produces a prototype-space decision as:

$$v_m = \arg\max_{c\in\{P,N\}} s_m^c. \tag{14}$$

The modality-wise decisions v_m form a voting process, where the majority vote defines the high-confidence sentiment direction. Disagreeing modalities are suppressed and replaced with prototype features consistent with the voted sentiment, preventing unreliable modalities from dominating fusion. The voting mechanism is activated only in the second stage, allowing the first stage to stabilize prototype formation before stronger decision-level consistency is enforced.

Decision-Level Alignment Loss. Decision-level alignment aims to construct a discriminative and cross-modally consistent sentiment prototype space, and to constrain multimodal fusion decisions to align with ground-truth sentiment labels. To this end, TLRA introduces three loss terms at the decision level: two for prototype structure learning and one for final sentiment prediction to further enhance robustness under modality missing and quality imbalance.

First, to ensure intrinsic discriminability within each modality, the positive (P) and negative (N) sentiment prototypes should be well separated in the prototype space. Let B_m^P and B_m^N denote the positive and negative prototypes of modality m. We define an intra-modality separation loss as:

$$\mathcal{L}_{\text{intra}} = \sum_{m\in\mathcal{M}} \max\big(0,\ \delta - \cos(B_m^P, B_m^N)\big), \tag{15}$$

where δ is a margin threshold enforcing clear class boundaries within each modality to enhance separability.

Second, to promote cross-modal consistency, prototypes corresponding to the same sentiment across different modalities are encouraged to align in the sentiment space. The cross-modal prototype alignment loss is defined as:

$$\mathcal{L}_{\text{inter}} = \sum_{\substack{m,n\in\mathcal{M}\\ m\neq n}} \Big(1-\cos(B_m^P, B_n^P) + 1 - \cos(B_m^N, B_n^N)\Big). \tag{16}$$

After prototype construction and alignment, the final sentiment prediction is obtained through the fusion module. A standard task supervision loss is applied:

$$\mathcal{L}_{\text{reg}} = \ell(\hat{y}, y), \tag{17}$$

where $\hat{y}$ and y denote the predicted and ground-truth sentiment labels, respectively, for each input sample. The overall decision-level optimization objective is:

$$\mathcal{L}_{\text{total}} = \lambda_3\mathcal{L}_{\text{intra}} + \lambda_4\mathcal{L}_{\text{inter}} + \lambda_5\mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{align}}. \tag{18}$$

Together, these losses enforce intra-modality discriminability, cross-modal consistency, and prediction correctness, enabling TLRA to produce robust sentiment predictions under modality quality imbalance. The loss weighting coefficients are selected via controlled validation, and the model remains stable across a reasonable range of settings, indicating that the performance gain primarily arises from the proposed representation- and decision-level alignment mechanisms rather than sensitive hyperparameter tuning.

4 Experiments

4.1 Experimental Setup, Datasets, and Metrics

Experimental Setup. We trained our model using PyTorch and 1 NVIDIA RTX 4090 GPU. Each modality encoder is initialized with a pretrained model, and alignment and fusion modules were learned end-to-end. Using AdamW with a learning rate of 1×10^{-4} and a cosine annealing schedule, the batch size is 16 and was trained over 100 epochs. All hyperparameters were kept constant across all experiments, where

		CMU-MOSI Dataset										CMU-MOSEI Dataset									
Set	Metric	CENET [2022]	TETFN [2022]	DMD [2023]	TFR-Net [2023]	ALMT [2023]	LNLN [2024]	UEFN [2025]	U-PNS [2025]	MECAM [2025]	Ours	CENET [2022]	TETFN [2022]	TFR-Net [2023]	ALMT [2023]	DiCMoR [2023]	IMDer [2023]	LNLN [2024]	U-PNS [2025]	MECAM [2025]	Ours
A	ACC	57.67	57.77	42.23	57.77	56.45	52.18	45.36	51.17	54.30	**62.50**	61.78	62.81	62.85	64.68	62.90	63.80	62.85	**69.39**	65.90	65.47
	F1	42.26	42.31	25.07	42.31	**67.09**	58.89	45.94	47.83	54.80	62.01	51.87	48.51	48.51	**71.71**	60.40	60.60	48.43	62.14	63.00	63.72
V	ACC	57.67	57.77	46.95	58.03	56.35	52.18	49.09	48.69	53.40	**62.65**	60.69	62.82	62.85	64.83	63.60	63.90	62.85	62.14	63.80	**66.04**
	F1	42.26	42.31	39.63	43.37	**66.89**	58.89	49.78	37.15	53.20	62.19	51.22	48.51	48.57	**71.79**	63.60	63.60	48.51	62.60	61.80	65.50
L	ACC	83.08	82.57	65.85	83.08	84.55	84.86	79.69	82.22	85.20	**85.24**	84.64	84.83	84.97	84.55	84.20	84.50	84.07	81.46	84.20	**85.09**
	F1	83.06	82.62	65.90	83.01	84.76	85.12	79.17	82.17	85.10	**85.17**	84.71	84.80	84.43	84.77	84.30	84.50	84.47	82.00	83.90	**85.05**
AV	ACC	57.67	57.77	42.23	57.82	56.40	52.18	51.78	55.98	56.60	**61.33**	63.23	62.82	62.85	54.28	65.20	64.90	62.85	**71.02**	66.60	66.61
	F1	42.26	42.31	25.07	42.43	**66.99**	58.89	50.34	56.09	55.80	61.33	52.41	48.53	48.88	**69.52**	64.40	63.50	48.43	60.97	63.90	65.58
AL	ACC	83.08	82.62	67.38	83.54	84.81	84.91	80.78	82.65	84.60	**85.83**	84.99	84.84	83.97	66.81	85.00	85.10	84.07	84.25	83.50	**85.86**
	F1	83.06	82.67	67.41	83.41	84.81	85.17	80.87	82.62	84.60	**85.97**	85.03	84.81	83.93	71.72	84.90	85.10	84.47	84.33	83.60	**85.76**
VL	ACC	83.08	82.52	67.38	82.72	84.60	84.86	82.99	82.94	84.10	**85.53**	**85.37**	85.08	84.95	67.13	84.90	85.00	84.39	85.32	83.00	84.81
	F1	83.06	82.57	67.51	82.71	84.81	85.12	82.64	82.90	84.10	**85.47**	**85.35**	85.05	84.69	71.76	84.90	85.00	84.71	85.34	83.00	84.81

Table 1: Quantitative comparison with representative state-of-the-art methods under missing-modality settings on the CMU-MOSI and CMU-MOSEI datasets. The best results in each column are highlighted in **bold**, and the second-best results are underlined. A, V, and L denote the audio, visual, and language modalities, respectively.

Set	Metric	ALMT [2023]	DMD [2023]	LNLN [2024]	DLF [2025]	UEFN [2025]	U-PNS [2025]	MECAM [2025]	M-BERT [2025]	Ours
CMU-MOSI	ACC	84.91	82.23	84.25	85.06	85.11	85.28	85.80	84.43	**86.28**
	F1	85.01	83.29	84.61	85.04	85.03	85.15	85.80	84.61	**86.24**
CMU-MOSEI	ACC	85.62	84.62	84.14	85.42	85.67	85.55	85.20	84.82	**85.88**
	F1	85.69	84.62	84.53	85.27	85.76	85.72	84.80	84.71	**85.86**

Table 2: Quantitative comparison with representative SOTA methods under full-modality settings on the CMU-MOSI and CMU-MOSEI datasets. The best results in each column are highlighted in **bold**, and the second-best results are underlined.

$\lambda_1 = \lambda_2 = 1.0$, $\lambda_3 = \lambda_4 = 0.7$, and $\lambda_5 = 0.2$. These coefficients are selected by one-factor-at-a-time validation to favor settings that remain stable across modality combinations rather than relying on exhaustive search. The additional cost of TLRA mainly comes from template retrieval and prototype matching, which scale linearly with the number of templates and prototypes. Since TLRA does not perform input-level modality reconstruction, the dominant computation remains in modality encoding, and the alignment modules introduce only limited overhead. This indicates that TLRA improves robustness without relying on expensive generation or reconstruction modules, making it more practical for deployment under resource-constrained settings.

Datasets. We assessed our methods against two publicly available benchmarks for multimodal sentiment analysis, CMU-MOSI [Zadeh *et al.*, 2016] and CMU-MOSEI [Zadeh *et al.*, 2018c]. For both datasets the videos and modal data were aligned; CMU-MOSI consists of 2199 utterances from 93 video clips, while CMU-MOSEI is a larger dataset containing 23,453 utterances from over 1000 speakers. With aligned multimodal annotations, we were able to perform end-to-end fusion learning with these datasets, as well as evaluate our algorithms under missing-modality settings.

Metrics. The performance metrics for both datasets were assessed using ACC and F1-score (binary sentiment), where samples were assigned to the negative class or the non-negative class based on standard thresholding of sentiment scores. ACC measures overall classification accuracy and F1-score complements ACC by balancing precision and recall.

4.2 Quantitative Evaluations

Missing-Modality Comparison. Table 1 shows a detailed quantitative comparison between our approach under missing-modality scenarios on both CMU-MOSI and CMU-MOSEI datasets, in addition to comparing other representative approaches for multimodal sentiment analysis under similar conditions of missing modalities, such as CENET [Wang *et al.*, 2022], TETFN [Wang *et al.*, 2023a], DMD [Li *et al.*, 2023], TFR-Net [Yuan *et al.*, 2021], ALMT [Zhang *et al.*, 2023], LNLN [Zhang *et al.*, 2024], UEFN [Wang *et al.*, 2025c], U-PNS [Chen *et al.*, 2025], MECAM [Wang *et al.*, 2025b], DiCMoR [Wang *et al.*, 2023b], and IMDer [Wang *et al.*, 2023c]. All methods are evaluated using ACC and F1-score across all conditions for binary sentiment classification.

Our approach performs better or remains competitive across most modality combinations, showing that TLRA mitigates degradation from incomplete multimodal information. The gains are clearer under severe missing settings, indicating stronger robustness with limited evidence. Representation-level alignment stabilizes incomplete features, while decision-level alignment reduces unreliable modality-wise predictions. Together, they address representation drift and decision inconsistency, producing more reliable sentiment predictions. Since missing-modality and reliability-imbalance settings both reflect modality reliability degradation, TLRA can handle unreliable evidence without relying on specific noise patterns, highlighting its adaptability to real-world conditions and varying data quality.

Full-Modality Comparison. Quantitative performance comparisons for our full-modality evaluation condition can be found in Table 2 where we compare our method to other representative multimodal sentiment analysis methods including ALMT, DMD, LNLN, DLF [Wang *et al.*, 2025a], UEFN, U-PNS and M-BERT [Rahman *et al.*, 2020].

Overall, the proposed framework achieves the highest ACC and F1-score on both CMU-MOSI and CMU-MOSEI, demonstrating robustness under complete-modality settings. Beyond performance gains, this suggests TLRA improves multimodal fusion even when all modalities are present. This also indicates TLRA is not limited to missing-modality sce-

No.	ME	RLA				DLA				CMU-MOSI Dataset L		VL		AVL	
	Dual	EncM	EncQ	Tpl	L_{align}	Stage1	Stage2	L_{inter}	L_{intra}	ACC	F1	ACC	F1	ACC	F1
0															
1	✓	✓								75.61	75.76	78.05	78.17	78.51	78.51
2	✓		✓							77.13	75.25	78.35	76.84	78.81	77.59
3	✓			✓						78.96	79.03	79.57	79.67	79.88	79.98
4	✓	✓	✓	✓						79.73	79.02	80.64	80.00	80.95	80.15
5	✓	✓	✓	✓	✓					80.03	80.15	81.40	81.42	81.71	81.82
6	✓	✓	✓	✓	✓	✓		✓	✓	83.62	83.64	83.78	84.15	84.22	84.14
7	✓	✓	✓	✓	✓		✓	✓	✓	83.32	83.31	83.47	83.56	83.53	83.63
8	✓	✓	✓	✓	✓	✓	✓			84.30	83.96	84.38	84.01	84.45	84.47
9	✓	✓	✓	✓	✓	✓	✓	✓		84.71	84.72	84.92	85.13	85.25	85.17
10		✓	✓	✓	✓	✓	✓	✓	✓	83.37	83.41	83.66	83.94	84.32	84.12
11	✓	✓	✓	✓	✓	✓	✓	✓	✓	**85.06**	**85.07**	**85.43**	**85.37**	**86.28**	**86.24**

No.0 Baseline **No.1-5** *Verify* Representation-Level Alignment **No.6-9** *Verify* Decision-Level Alignment **No.10** *Verify* Modality Encoding **No.11** Our Method

ME Modality Encoding **RLA** Representation-Level Alignment **DLA** Decision-Level Alignment **Dual** Dual-Path
EncM Modal Encoding **EncQ** Query Encoding **Tpl** Template L_{align} Representation Alignment Loss
Stage1 Soft Prototype Guidance **Stage2** Prototype Voting and Suppression L_{inter} Cross-Modal Prototype Alignment Loss L_{intra} Intra-Modality Separation Loss

Table 3: Ablation study of Modality Encoding, Representation-Level Alignment, and Decision-Level Alignment components on the CMU-MOSI dataset under different modality settings.

narios; its decision-level reference mechanism improves fusion reliability when modalities exhibit inconsistent quality or conflicting sentiment cues. This further highlights the effectiveness of reference-guided decision alignment in promoting consistent and reliable multimodal predictions.

4.3 Ablation Experiments

We conduct systematic ablation studies on CMU-MOSI, with results summarized in Table 3. By progressively adding modality encoding (ME), representation-level alignment (RLA), and decision-level alignment (DLA) to a baseline, we isolate the contribution of each component.

Results from No. 1 to No. 5 show that RLA consistently improves performance, indicating that stable references effectively mitigate representation drift under modality missing. Results from No. 6 to No. 9 further confirm the effectiveness of DLA; notably, the two-stage strategy yields consistent gains by suppressing unreliable modalities via prototype guidance and voting to enhance cross-modal consistency.

The full model (No. 11) achieves the best performance across all modality settings, demonstrating that jointly aligning representations and decisions is critical for robustness under modality-missing conditions.

4.4 Cross-Dataset Generalization

Table 4 reports the cross-dataset evaluation results, where models are trained on one dataset and tested on the other. This setting is particularly challenging because speakers, topics, and expression styles vary significantly across datasets, often leading to overfitting on dataset-specific patterns. Compared with representative baselines, our method consistently achieves the best performance and exhibits smaller degradation when transferring across datasets. This indicates that TLRA effectively reduces reliance on dataset-specific cues and captures more intrinsic and transferable sentiment structures. As a result, TLRA learns more stable and dataset-invariant sentiment representations, enabling robust generalization under distribution shifts and unseen data conditions.

4.5 Prototype Similarity Analysis

Figure 4 visualizes similarity differences between selected samples and their prototypes across training stages. We select ten positive and ten negative samples and evaluate them at 10, 40, and 70 epochs. In early stages, high contrast indicates modality-specific differences, suggesting that the feature space is dominated by modality variations and inconsistent sentiment representations. As training proceeds, the colors become lighter and more uniform, showing that samples align with their prototypes and converge toward a shared sentiment structure. At later stages, samples show consistent prototype similarity, indicating that the prototype bank has become stable class-level anchors. This trend supports the two-stage strategy, where soft guidance stabilizes the prototype space before voting-based suppression is applied.

Model	Training on CMU-MOSI		Training on CMU-MOSEI		Training on CMU-MOSEI		Training on CMU-MOSI	
	Testing on **CMU-MOSI** (ACC/F1)				Testing on **CMU-MOSEI** (ACC/F1)			
MAG-BERT [2020]	84.43	84.61	81.40	81.64	84.82	84.71	77.04	77.16
DMD [2023]	83.23	83.29	61.74	52.35	84.62	84.62	37.15	20.12
ALMT [2023]	84.91	85.01	79.45	79.32	85.62	85.69	70.02	71.40
LNLN [2024]	84.25	84.61	76.03	71.74	84.14	84.53	69.04	66.99
DLF [2025]	85.06	85.04	80.18	80.12	85.42	85.27	74.77	75.15
Ours	**86.28**	**86.24**	**82.16**	**82.08**	**85.88**	**85.86**	**79.91**	**80.11**

Table 4: Cross-dataset evaluation results on CMU-MOSI and CMU-MOSEI, where models are trained on one dataset and tested on the other. This setting evaluates generalization under mismatch.

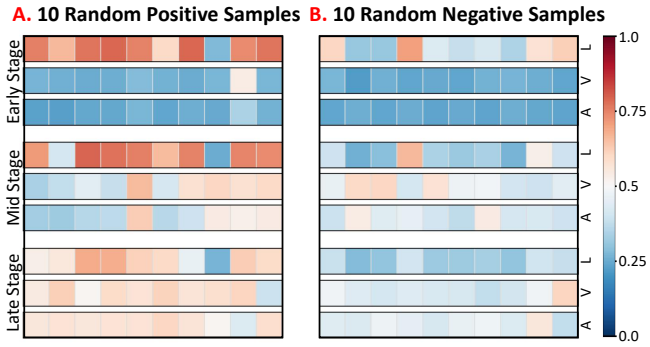

Figure 4: Similarity difference visualization between randomly sampled sentiment samples and sentiment prototypes across training stages. Lighter colors indicate improved cross-modal alignment and more consistent sentiment representations.

5 Conclusion

We present the Two-Level Reference Alignment framework to address decision drift in multimodal sentiment analysis under missing or unreliable modality conditions. TLRA aligns feature representations and final decisions to stable references, mitigating representation deviation and biased fusion caused by incomplete or unreliable evidence. Experiments on CMU-MOSI and CMU-MOSEI show robust performance across missing-modality, full-modality, and cross-dataset settings. These results demonstrate that feature- and decision-level reference constraints improve stability, robustness, and generalization under diverse modality degradation scenarios, while reducing sensitivity to dataset-specific sentiment patterns and modality inconsistencies. The framework is general and can be extended to other multimodal learning tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62172246), the Excellent Young Scientists Fund of the Natural Science Foundation of Shandong Province (No. ZR2024YQ071), the Key Laboratory of Forensic Examination for Sichuan Provincial Universities (No. 2024YB01), and the Fundamental Research Funds for the Central Universities (No. 22CX06037A).

References

[Arevalo *et al.*, 2017] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[Baltrušaitis *et al.*, 2016] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

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

[Dai *et al.*, 2025] Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24507–24517, 2025.

[Devlin *et al.*, 2019] 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, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Guo *et al.*, 2024] Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. *arXiv preprint arXiv:2407.05374*, 2024.

[Han *et al.*, 2021] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*, 2021.

[Hazarika *et al.*, 2020] 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*, pages 1122–1131, 2020.

[Li *et al.*, 2023] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6631–6640, 2023.

[Li *et al.*, 2024] Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, et al. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. *Advances in Neural Information Processing Systems*, 37:28515–28536, 2024.

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

[Liu *et al.*, 2018] 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, 2018.

[Liu *et al.*, 2024] Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 101:101973, 2024.

[Liu, 2022] Bing Liu. *Sentiment analysis and opinion mining*. Springer Nature, 2022.

[Mao *et al.*, 2023] Huisheng Mao, Baozheng Zhang, Hua Xu, Ziqi Yuan, and Yihe Liu. Robust-msa: Understanding the impact of modality noise on multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16458–16460, 2023.

[McFee *et al.*, 2015] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. *SciPy*, 2015:18–24, 2015.

[Neverova *et al.*, 2015] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[Pham *et al.*, 2019] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6892–6899, 2019.

[Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.

[Rahman *et al.*, 2020] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating

multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2359–2369, 2020.

[Reza *et al.*, 2024] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Tsai *et al.*, 2019] 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 conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

[Wang *et al.*, 2019] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7216–7223, 2019.

[Wang *et al.*, 2022] 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*, 25:4909–4921, 2022.

[Wang *et al.*, 2023a] 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.

[Wang *et al.*, 2023b] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034, 2023.

[Wang *et al.*, 2023c] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

[Wang *et al.*, 2025a] Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188, 2025.

[Wang *et al.*, 2025b] Rui Wang, Yaoyang Wang, Erik Cambria, Xuhui Fan, Xiaohan Yu, Yao Huang, Xiaosong E, and Xianxun Zhu. Contrastive-based removal of negative information in multimodal emotion analysis. *Cognitive Computation*, 17(3):107, 2025.

[Wang *et al.*, 2025c] Shuai Wang, K Ratnavelu, and Abdul Samad Bin Shibghatullah. Uefn: Efficient uncertainty estimation fusion network for reliable multimodal sentiment analysis. *Applied Intelligence*, 55(3):171, 2025.

[Wu *et al.*, 2024] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

[Xie *et al.*, 2024] Zhuyang Xie, Yan Yang, Jie Wang, Xiaorong Liu, and Xiaofan Li. Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7657–7670, 2024.

[Yu *et al.*, 2021] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 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, 2021.

[Yu *et al.*, 2024] Xiaomin Yu, Feiyang Wang, and Ziyue Qiao. Spikemo: Enhancing emotion recognition with spiking temporal dynamics in conversations. *arXiv preprint arXiv:2411.13917*, 2024.

[Yuan *et al.*, 2021] 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.

[Zadeh *et al.*, 2016] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[Zadeh *et al.*, 2018a] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Zadeh *et al.*, 2018b] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Zadeh *et al.*, 2018c] 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, 2018.

[Zeng *et al.*, 2022] Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 2924–2934, 2022.

[Zhang *et al.*, 2023] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*, 2023.

[Zhang *et al.*, 2024] Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. *Advances in Neural Information Processing Systems*, 37:55943–55974, 2024.