Recovering Coherent Affective Patterns: Addressing Modality Missing in Multimodal Sentiment Analysis

Huiting Huang[1,2], Tieliang Gong[1,2*], Kai He[3], Wen Wen[1,2], Weizhan Zhang[1,2], Mengling Feng[3]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
[2]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University, Xi'an, China
[3]Saw Swee Hock School of Public Health, National University of Singapore, Singapore
{huiting.huang}@stu.xjtu.edu.cn, {gongtl, zhangwzh}@xjtu.edu.cn, {kai_he, ephfm}@nus.edu.sg, wen190329@gmail.com

Abstract

Multimodal sentiment analysis (MSA) seeks to decode human emotions by integrating heterogeneous modalities. However, real-world scenarios often involve missing or misaligned data due to sensor failures or transmission errors, leading to disrupted temporal dynamics and degraded cross-modal correlations. To address these challenges, we propose RECAP (REcovery of Coherent Affective Patterns), a robust two-stage framework to restore temporal and structural emotional integrity under modality incompleteness. The first stage employs a causality-aware adversarial generator for multi-granularity temporal reconstruction, complemented by a contrastive mutual information factorization module that disentangles shared and modality-specific semantics. The second stage introduces a mutual information-guided attention fusion mechanism with a ranking-based objective, enabling adaptive integration of complementary signals for refined prediction. Extensive experiments on MOSI, MOSEI, and SIMS under various missing-modality conditions demonstrate that RECAP consistently outperforms state-of-the-art methods. Notably, it improves ACC-7 on MOSI by 2.71 percentage points and F1 on SIMS by 6.38 percentage points. These results verify the performance of RECAP in terms of capturing fine-grained emotional cues and robustness.

Code — https://github.com/Taylor-HHT/RECAP-MSA

Introduction

Multimodal sentiment analysis (MSA) integrates complementary signals from diverse sensory modalities to achieve comprehensive understanding of human emotions. It has demonstrated remarkable success in a multitude of domains including affective computing (Yi et al. 2024), healthcare (Yao et al. 2024; Lan et al. 2025), social media understanding (Deng, Ananthram, and McKeown 2025), and human-computer interaction (Jiang et al. 2020). However, existing MSA methods typically depend on the unrealistic assumption that all modalities are fully observed and perfectly aligned (Zhang et al. 2023; Yi et al. 2024; He et al. 2025), which compromises their effectiveness in modality absence settings. In the real world, modality inputs are frequently incomplete or degraded due to sensor malfunctions, asynchronous sampling, or environmental noise (Yao et al. 2024; Sun et al. 2024). Such imperfections pose significant challenges to model robustness and underscore the necessity for models capable of inferring meaningful semantics from partial and misaligned multimodal data (Zhang et al. 2024).

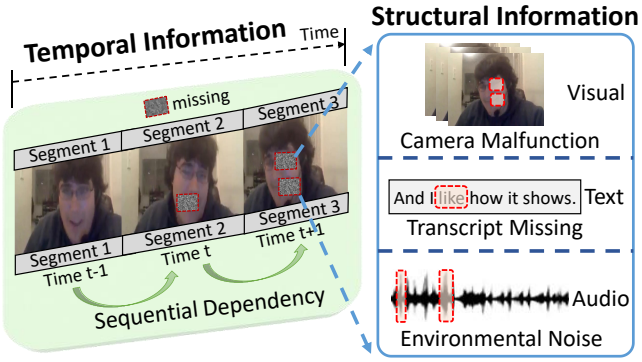

Figure 1: Illustration of the motivation. Left: Temporal continuity across frames supports missing segment inference. Right: Modality information may be degraded due to occlusion, transcript loss, or noise, requiring cross-modal consistency and complementary cues for recovery.

Recent effort has been dedicated to addressing incomplete modality data by leveraging joint learning (Tang et al. 2021; Zeng, Liu, and Zhou 2022; Li, Yang, and Zhang 2023) and generative methods (Sun et al. 2023; Zhang, Wang, and Yu 2024). Joint learning approaches aim to infer the latent representations of missing modalities from available ones. For instance, ShaSpec (Wang et al. 2023b) addresses missing modality inference by decomposing observed inputs into shared and modality-specific components, leveraging the shared space to approximate the missing representations. While attractive, its decomposition relies on auxiliary objectives like domain classification and distribution alignment, which offers only indirect control over the semantic disentanglement. Moreover, directly utilizing multimodal data with uncertain missing information often leads to poor common space projection, which consequently degrades overall performance (Sun et al. 2024). Additionally, the aforementioned work is developed to handle the total of one or more modalities, which seldom reflects real-world

*Corresponding author.

scenarios where stochastic and partial missing data across modalities is far more common (Tao et al. 2025).

The other representative method is generative learning, which generate missing data based on observed modality distributions. For example, LNLN (Zhang, Wang, and Yu 2024) designs a language-dominant framework that estimates the completeness of the language modality and generates proxy features from visual and acoustic inputs to compensate language degradation. While effective in certain settings, these methods often overlook the inherent temporal causality and structural affective patterns embedded in multimodal data. As illustrated in Figure 1, multimodal videos encode affective information along two essential dimensions: *temporal dynamics* and *structural modality interactions*. Temporally, video frames exhibit sequential and causal dependencies, where each frame builds on prior context, thereby providing an inductive bias for inferring missing segments. Structurally, sentiment is distributed across heterogeneous modalities such as text, visual and audio, each capturing distinct yet complementary affective cues. Disruptions in any modality can compromise the affective interplay, making it imperative for generative models to account for both cross-modal consistency and modality complementarity during the reconstruction process.

To bridge this gap, we propose RECAP, a novel two-stage framework for **RE**covering **C**oherent **A**ffective **P**atterns. Unlike prior approaches, RECAP leverages a generative paradigm to explicitly reconstruct coherent affective dynamics that are both temporally grounded and structurally complete. Specifically, RECAP integrates a Hierarchical Causality-aware Adversarial Generation (H-CAG) mechanism and a contrastive Factorized Information Decomposition (FID) strategy, which together enable robust recovery by modeling temporal dependencies and quantifying structured representations. The Mutual Information-guided Ranking (MIR) module further facilitates adaptive signal fusion by prioritizing informative modalities according to their task-relevant informativeness. Moreover, our method is generally applicable to common multimodal scenarios encompassing stochastic and partial multimodal inputs.

Overall, our contributions are summarized as follows:

- We propose a novel two-stage framework RECAP, which consists of modality completion and adaptive fusion, enabling the model to restore coherent temporal and structural affective cues while effectively integrating complementary information for robust MSA. To the best of our knowledge, RECAP is the first to explicitly model both temporal and structural affective information caused by modality incompleteness.
- Our framework integrates two key modules H-CAG and FID, to ensure high-fidelity modality completion. Specifically, H-CAG leverages self-supervised adversarial learning over multi-granularity temporal segments to infer missing modality features while preserving causal continuity across time. FID promotes structural integrity by disentangling shared and unique emotional cues via mutual information optimization.
- Extensive experiments on benchmark datasets show that RECAP achieves state-of-the-art performance and maintains relatively stable accuracy across various missing rates. Our method also demonstrates superior capability in fine-grained sentiment classification by capturing subtle emotional cues and modeling a coherent affective space, even under partial observations. Visualization results further confirm that the reconstructed modalities closely align with the real distribution.

Methodology

Overview

Figure 2 illustrates the overall architecture of RECAP designed to handle partial modality incompleteness in MSA. It begins by extracting raw features and simulating real-world missing modality scenarios. The specific embedding modules first transform heterogeneous inputs into a more consistent representation, standardizing the dimensionality. The framework then proceeds in two stages: (a) modality completion, and (b) fusion and prediction. In the first stage, missing modality features are reconstructed by capturing both temporal dependencies and structural semantics. This process is driven by H-CAG and FID. H-CAG learns temporal and intra-modal patterns for generative recovery, while FID serves as an auxiliary constraint on H-CAG, refining its generative process by promoting inter-modal representational disentanglement. Moreover, a reconstructor maps the latent outputs to complete modality representations, bridging the gap between generation and prediction. In the second stage, the MIR module adaptively fuses the generator's outputs into a unified representation, which is then fed into a classifier to predict the final sentiment label.

Multimodal Missing Input

We focus on the multimodal video dataset comprising three modalities: language (l), audio (a) and visual (v), each represented as a time-series. Following the previous work (Zhang et al. 2023), each raw modality input is processed using standard toolkits to obtain sequential features $X_m \in \mathbb{R}^{T_m \times d_m}$, $m \in \{l, a, v\}$, where T_m is the sequence length, and d_m is the feature dimension. We then simulate corrupted multimodal inputs $\tilde{X}_m \in \mathbb{R}^{T_m \times d_m}$ by randomly omitting portions of the data X_m, aligned with Zhang, Wang, and Yu (2024) to ensure fair and consistent evaluation. In particular, the corrupted visual and acoustic modalities are obtained by replacing missing values with zeros, whereas the language modality is corrupted by substituting erased tokens with the [UNK] token as used in BERT (Devlin et al. 2019). Given the corrupted multimodal input $\tilde{X}_m$, the objective is to reconstruct and integrate the missing modality information to accurately predict the target sentiment label y.

Modality Completion Stage

In modality completion stage, we recover missing features by exploiting both temporal causality and structural patterns. Specifically, we introduce two complementary modules: Hierarchical Causality-aware Generation (H-CAG), which captures temporal dynamics across local and global granularities, and Factorized Information Decomposition (FID),

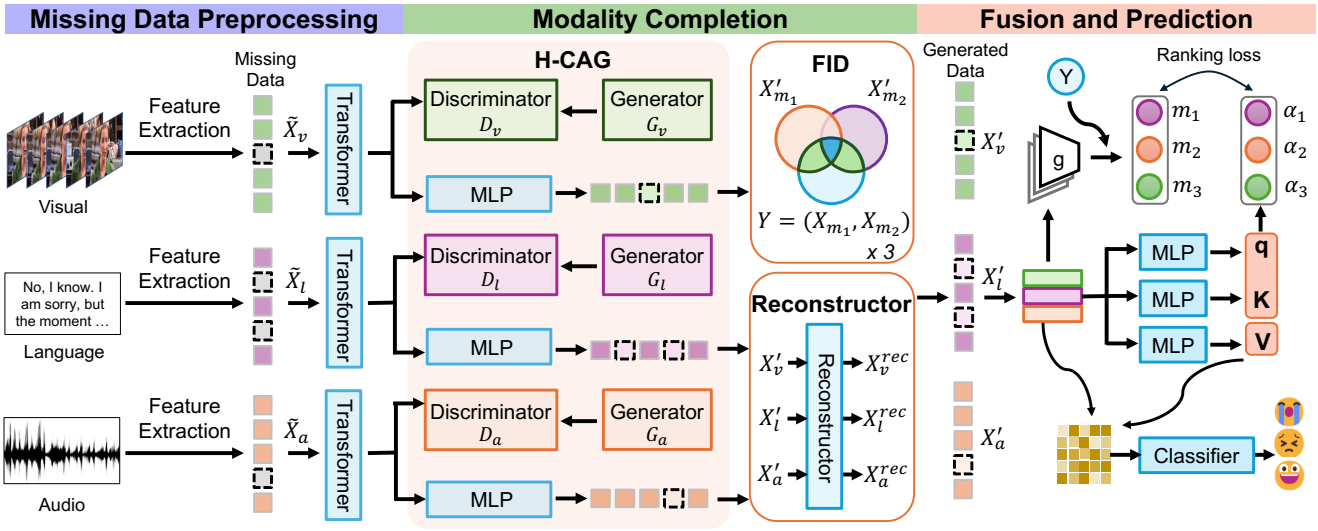

Figure 2: The overall architecture of the proposed framework RECAP.

which learns to separate shared and unique signals through mutual information factorization. Together, H-CAG and FID enhance the temporal continuity and semantic integrity of corrupted inputs. Additionally, a reconstruction loss is applied to reinforce the recovery process.

Hierarchical Causality-aware Adversarial Generation (H-CAG). We introduce multi-granularity approach to hierarchically learn dependencies at both local (short-term) and global (long-term) temporal scales. Given the corrupted input sequence $\tilde{X}_m$, we first divide it into non-overlapping temporal segments at multiple granularities, denoted as $S_m^{(k)} = \{s_{m,1}^{(k)}, s_{m,2}^{(k)}, \ldots, s_{m,N_k}^{(k)}\}$, where N_k is the number of segments at the k-th granularity. Here, granularity denotes the scale of temporal segmentation. Lower (coarse) granularity captures global affective trends via longer segments, whereas higher (fine) granularity retains subtle emotional shifts through shorter segments. For each segment pair $(s_{m,t}^{(k)}, s_{m,t+1}^{(k)})$, the generator G_m then takes the earlier segment as input and generates the next:

$$\hat{s}_{m,t+1}^{(k)} = G_m(s_{m,t}^{(k)}). \tag{1}$$

To encourage semantic plausibility and temporal causality, the generator is trained in an adversarial manner against the discriminators D_m, which aim to distinguish real future segments $s_{m,t+1}^{(k)}$ from generated ones $\hat{s}_{m,t+1}^{(k)}$. The generators G_m, in turn, seek to fool discriminators into classifying the generated outputs as real sequences.

To further emphasize high-correlation segment pairs, we introduce a cosine similarity weight $\omega_t^{(k)} \in [0, 1]$, computed between $s_{m,t}^{(k)}$ and $s_{m,t+1}^{(k)}$:

$$\omega_t^{(k)} = \left(1 + \cos(s_{m,t}^{(k)}, s_{m,t+1}^{(k)})\right)/2. \tag{2}$$

The weights encourage precise reconstruction of highly correlated transitions and are used to modulate the generator's loss, which is defined as follows:

$$\mathcal{L}_{\text{gen}}^{(k)} = \omega_t^{(k)} \cdot \mathcal{L}_{\text{BCE}}(D_m(\hat{s}_{m,t+1}^{(k)}), 1), \tag{3}$$

where the binary cross-entropy (BCE) loss is given by:

$$\mathcal{L}_{\text{BCE}}(p, z) = -\left[z \cdot \log(p) + (1 - z) \cdot \log(1 - p)\right], \tag{4}$$

with prediction value $p \in [0, 1]$ and target label $z \in \{0, 1\}$.

We further employ the following loss function to optimize the discriminator for accurate prediction:

$$\mathcal{L}_{\text{disc}}^{(k)} = \frac{1}{2}\left[\mathcal{L}_{\text{BCE}}(D_m(s_{m,t+1}^{(k)}), 1) + \mathcal{L}_{\text{BCE}}(D_m(\hat{s}_{m,t+1}^{(k)}), 0)\right] \tag{5}$$

comprising real-vs-fake classification for both authentic and generated samples.

We compute the adversarial loss over a set of temporal granularities $\mathcal{G}$, and the final loss for modality m is:

$$\mathcal{L}_{\text{adv}}^{m} = \sum_{k \in \mathcal{G}} \left(\mathcal{L}_{\text{gen}}^{(k)} + \mathcal{L}_{\text{disc}}^{(k)}\right). \tag{6}$$

The total loss of H-CAG module across all modalities is a weighted combination:

$$\mathcal{L}_{\text{H-CAG}} = w_l \cdot \mathcal{L}_{\text{adv}}^{l} + w_a \cdot \mathcal{L}_{\text{adv}}^{a} + w_v \cdot \mathcal{L}_{\text{adv}}^{v}, \tag{7}$$

where w_l, w_a, w_v are modality-specific weights.

The trained generators G_m yield the enhanced modality representations X'_m, which serve as inputs to both the FID module and the subsequent fusion and prediction stage.

Factorized Information Decomposition (FID). We propose the Factorized Information Decomposition (FID) module, inspired by FactorCL (Liang et al. 2023), which utilizes external annotations to extract task-relevant signals for classification and retrieval. In contrast to this method, FID operates in a fully self-supervised manner and leverages the inherent structure of multimodal data as implicit supervision to enable the extraction of complete-relevant features,

making it more generalizable and better suited for modality recovery under incomplete MSA conditions. More specifically, FID structurally disentangles shared and modality-specific representations by minimizing upper bounds of mutual information to suppress redundancy, while maximizing lower bounds to preserve informative content.

We present the process of decomposing modality representations into complete-relevant information in Figure 3. Each enhanced feature X'_m obtained from the H-CAG module, is considered to preserve the core semantics of its complete counterpart X_m, satisfying $X'_m \sim p(X'_m|X_m)$. We further treat the complete features X_m as a surrogate task signal Y, guiding the extraction of modality-invariant signals and complementary modality-specific cues. Formally, the learning objective involves mutual information calculation. For two arbitrary modalities m_1 and m_2, $S = I(X'_{m_1}; X'_{m_2}; Y)$ captures the shared semantics aligned with complete signal $Y = (X_{m_1}, X_{m_2})$, while $U_{m_1} = I(X'_{m_1}; Y|X'_{m_2})$ and $U_{m_2} = I(X'_{m_2}; Y|X'_{m_1})$ quantifying residual information unique to each modality.

We adopt a dual-bound strategy to estimate the MI terms in a tractable and scalable fashion, leveraging scalable lower and upper bounds. Specifically, the shared component loss of modality m_1 and m_2 is defined as:

$$\mathcal{L}_{\text{shared}}^{(m_1,m_2)} = -I_{NCE}(X'_{m_1}; X'_{m_2}) + I_{NCE-CLUB}(X'_{m_1}; X'_{m_2}|X_{m_1}, X_{m_2}). \quad (8)$$

Here, the first term applies the InfoNCE lower bound (Oord, Li, and Vinyals 2018) to preserve inter-modal commonality by encouraging semantic similarity between positive pairs (x, x^+) and contrasting them against negative samples x^-. The second term adopts a CLUB-based (Cheng et al. 2020) conditional upper bound to penalize redundant information retained from the original inputs and guide the learning of cleaner, modality-specific features, which is defined as:

$$I_{NCE-CLUB}\left(X'_{m_1}; X'_{m_2}|X_{m_1}, X_{m_2}\right) = \mathbb{E}\Big[\mathbb{E}\left[f^*\left(x'_{m_1}, x'^{+}_{m_2}, x_{m_1}, x_{m_2}\right)\right] - \mathbb{E}\left[f^*\left(x'_{m_1}, x'^{-}_{m_2}, x_{m_1}, x_{m_2}\right)\right]\Big], \quad (9)$$

where f^* is a learnable critic distinguishing positive and negative modality pairs.

In parallel, the unique loss for modality m_1 conditioned on its paired modality m_2, is formulated as:

$$\mathcal{L}_{\text{unique}}^{m_1} = -I_{NCE}(X_{m_1}; X'_{m_1}) + I_{NCE-CLUB}(X'_{m_1}; X'_{m_2}) - I_{NCE}(X'_{m_1}; X'_{m_2}|X_{m_1}, X_{m_2}). \quad (10)$$

Finally, FID applies the shared and unique objectives over all modality pairs (l, a), (l, v), and (a, v). The overall factorization loss is defined as:

$$\mathcal{L}_{\text{FID}} = \sum_{\substack{(m_1,m_2)\in \\ \{(l,a),(l,v),(a,v)\}}} \left(\mathcal{L}_{\text{shared}}^{(m_1,m_2)} + \mathcal{L}_{\text{unique}}^{m_1} + \mathcal{L}_{\text{unique}}^{m_2}\right). \quad (11)$$

The above objective empowers FID module to explicitly regulate the flow of information between complete and enhanced modalities, ensuring that reconstructions reflect both shared semantics and modality-specific signals. As a result, the proposed method preserves the inherent consistency and complementarity across modalities, thereby maintaining the structural integrity of affective representations.

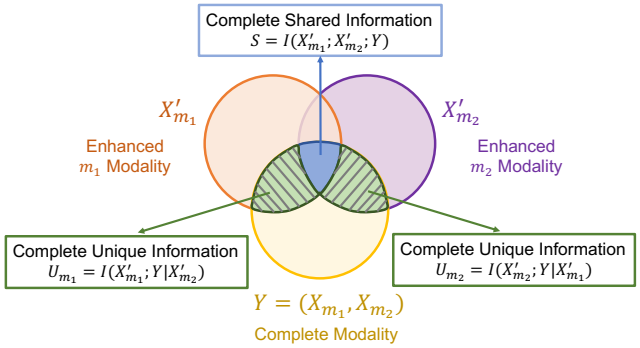

Figure 3: Illustration of the FID module. Enhanced modalities (e.g., X'_{m_1}, X'_{m_2}) are guided by their corresponding complete versions (X_{m_1}, X_{m_2}) to recover both shared and unique information through mutual information estimation.

Reconstructor. To ensure that the generated representations preserve the essential information of the original input, we introduce an auxiliary reconstructor E_{rec}, composed of two Transformer layers, which explicitly encourages the generated inputs X'_m to approximate X_m by reconstructing the original modality features. The reconstructor takes X'_m as input and outputs:

$$X_m^{rec} = E_{\text{rec}}(X'_m). \quad (12)$$

We apply a mean squared error loss across all modalities:

$$\mathcal{L}_{\text{rec}} = \sum_{m\in\{l,a,v\}} \|X_m^{rec} - X_m\|_2^2. \quad (13)$$

The above objective explicitly constrains the reconstructed features to align closely with the original inputs, thereby enhancing semantic fidelity.

Completion Training Objective. In the completion stage, we jointly optimize three objectives:

$$\mathcal{L}_{\text{stage1}} = \lambda_{\text{adv}}\mathcal{L}_{\text{H-CAG}} + \lambda_{\text{fid}}\mathcal{L}_{\text{FID}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}, \quad (14)$$

where coefficients $\lambda_{\text{adv}}, \lambda_{\text{fid}}, \lambda_{\text{rec}}$ balance temporal consistency, structural disentanglement, and feature-level reconstruction. As the training relies only on uncorrupted inputs and requires no task-specific labels, the learned generators generalize well across downstream tasks (Li, Savarese, and Hoi 2022; Grau et al. 2023).

Fusion and Prediction Stage

The fusion and prediction stage aims to adaptively integrate multimodal information and accurately predict sentiment. By leveraging the resulting modality-complete representations in previous stage, we obtain enhanced features for each modality, which are then integrated through a novel Mutual Information-guided Ranking mechanism (MIR) that assigns importance scores regarding informativeness to each modality. The fused and unified representation is passed to a task-specific prediction head to generate the final output.

Mutual Information-guided Ranking Fusion (MIR). To explicitly align the learned attention weights with modality-specific informativeness contribution, we introduce a ranking loss supervised by the mutual information scores.

Given the reconstructed features from the generator X'_m, each modality stream produces a prediction $\hat{Y}_m$ using a modality-specific predictor g_m, and its corresponding MI score is approximated using negative mean squared error with respect to the ground-truth label Y, which serves as a widely adopted and empirically stable proxy for mutual information in prior work (Achille and Soatto 2018; Amjad and Geiger 2019). These MI scores indicate the predictive reliability of each modality.

Let $\mathcal{L}_{\text{MI}}^m$ denote the MI-based auxiliary loss for modality. The MI-guided attention ranking loss is then formulated as:

$$\mathcal{L}_{\text{rank}} = \frac{1}{2} \sum_{i \neq j} \mathbb{I}\left[\mathcal{L}_{\text{MI}}^i < \mathcal{L}_{\text{MI}}^j\right] \cdot \max\left(0, \alpha_j - \alpha_i + \epsilon\right), \quad (15)$$

where α_m is the attention weight assigned to modality m, ϵ is a margin hyperparameter, and $\mathbb{I}[\cdot]$ is the indicator function. Equation 15 penalizes situations where a modality with stronger predictive performance (i.e., lower MI loss) is assigned a lower attention weight, thereby enforcing consistency between informativeness and attention importance.

In parallel, we project modality-specific features into key-query-value (KQV) triplets and compute attention logits via:

$$\text{logit}_m = \frac{1}{\sqrt{d}} \langle \mu_q^{(m)}, \mu_k^{(m)} \rangle, \quad (16)$$

where $\mu_q^{(m)}$ and $\mu_k^{(m)}$ are temporal means of the query and key vectors of modality m, with scaling dimension d. The final fused feature is a weighted aggregation over values:

$$\mathbf{f}_{\text{fused}} = \sum_{m \in \{l,a,v\}} \alpha_m \cdot \mu_v^{(m)} \in \mathbb{R}^d, \quad (17)$$

Task Prediction. The fused representation $\mathbf{f}_{\text{fused}}$ is passed through a final prediction head consisting of a linear projection layer. The task loss $\mathcal{L}_{\text{task}}$ is defined as:

$$\mathcal{L}_{\text{task}} = \begin{cases} \|y - \hat{y}\|_2^2, & \text{for regression} \\ -\sum_{c=1}^{C} y_c \log \hat{y}_c, & \text{for classification} \end{cases} \quad (18)$$

where y is the ground-truth label and $\hat{y}$ is the model prediction. For classification, y_c and $\hat{y}_c$ denote the true label and predicted probability of class c, respectively. The overall training objective of the fusion and prediction stage is:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{task}} \mathcal{L}_{\text{task}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}, \quad (19)$$

which combines the task objective with a ranking alignment loss, encouraging both performance consistency and interpretability in the fusion process.

Experiments

Experimental Settings

Datasets. We assess the performance of our method on popular multimodal sentiment analysis benchmark datasets, including CMU-MOSI (Zadeh et al. 2016), CMU-MOSEI (Zadeh et al. 2018) and CH-SIMS (Yu et al. 2020) datasets. Serving as standard benchmarks, these datasets provide a diverse testbed for assessing generalization. CMU-MOSI and CMU-MOSEI include spontaneous English content, while CH-SIMS introduces Chinese data for cross-lingual evaluation. All three are multimodal video datasets containing textual, visual, and acoustic modalities.

Evaluation Settings. To evaluate robustness under modality degradation, we conduct experiments with missing rates $r \in \{0.0, 0.1, \ldots, 0.9\}$, randomly masking a proportion r of features in each modality at test time. For example, $r = 0.5$ implies 50% of each modality is omitted. Unlike (Yuan et al. 2021), we exclude $r = 1.0$, as complete erasure across all modalities provides no informative signal. Final results in Table 1 and 2 are averaged over all rates to reflect overall performance under varying levels of input sparsity.

Evaluation Metrics. Following prior work (Yu et al. 2021; Zhang et al. 2023), we evaluate our model under both classification and regression settings. For classification, we report weighted F1 and binary accuracy (Acc-2). On MOSI and MOSEI, Acc-2 and F1 are evaluated under two protocols: negative vs. positive (left-side value of "/"), and negative vs. non-negative (including label 0, right-side value of "/"). We also report 5-class (Acc-5), 7-class (Acc-7) accuracy for MOSI and MOSEI, and Acc-2, F1, 3-class (Acc-3), Acc-5 for CH-SIMS. For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). In all metrics except MAE, higher values indicate better performance.

Implementation Details. We train all models for 200 epochs using Python 3.9.18, PyTorch 2.2.2, and CUDA 12.2 on NVIDIA RTX 4090 GPUs with 24GB memory. The operating system is Ubuntu 24.04. The AdamW optimizer is adopted with a learning rate of 1e-4 and weight decay of 1e-4. The batch size is set to 64 for all datasets. For MOSI and SIMS, the generators, discriminators, reconstructors, and fusion modules are implemented as Transformer architectures with 2 layers, 8 attention heads, and a hidden size of 128. For MOSEI, the generator adopts a deeper architecture initialized from the pre-trained ImageBind (Girdhar et al. 2023) model to better accommodate the dataset's greater volume and complexity, while the other modules follow the same Transformer configuration as in MOSI and SIMS.

Baseline Models

We benchmark RECAP against eleven state-of-the-art MSA approaches across all datasets, including MISA (Hazarika, Zimmermann, and Poria 2020), Self-MM (Yu et al. 2021), MMIM (Han, Chen, and Poria 2021), TFR-Net (Yuan et al. 2021), CENET (Wang et al. 2022), TETFN (Wang et al. 2023a), ALMT (Zhang et al. 2023), BI-Mamba (Yang et al. 2024), LNLN (Zhang, Wang, and Yu 2024), MASCF (Chen, Tang, and Liu 2025), TF-Mamba (Li et al. 2025).

Results and Analysis

Comparison to State-of-the-art Methods

Table 1 and Table 2 summarize the average performance of RECAP and eleven strong baselines across three datasets un-

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
MISA 2020	29.85	33.08	71.49 / 70.33	71.28 / 70.00	1.085	0.524	40.84	39.39	71.27 / 75.82	63.85 / 68.73	0.780	0.503
Self-MM 2021	29.55	34.67	70.51 / 69.26	66.60 / 67.54	1.070	0.512	44.70	45.38	73.89 / 77.42	68.92 / 72.31	0.695	0.498
MMIM 2021	31.30	33.77	69.14 / 67.06	66.65 / 64.04	1.077	0.507	40.75	41.74	73.32 / 75.89	68.72 / 70.32	0.739	0.489
TFR-Net 2021	29.54	34.67	68.15 / 66.35	61.73 / 60.06	1.200	0.459	46.83	34.67	73.62 / 77.23	68.80 / 71.99	0.697	0.489
CENET 2022	30.38	37.25	71.46 / 67.73	68.41 / 64.85	1.080	0.504	47.18	47.83	74.67 / 77.34	70.68 / 74.08	0.685	0.535
TETFN 2023	30.30	34.34	69.76 / 67.68	65.69 / 63.29	1.087	0.507	30.30	47.70	69.76 / 67.68	65.69 / 63.29	1.087	0.508
ALMT 2023	30.30	33.42	70.40 / 68.39	72.57 / 71.80	1.083	0.498	40.92	41.64	76.64 / 77.54	77.14 / 78.03	0.674	0.481
BI-Mamba 2024	31.20	34.02	71.74 / 71.12	71.83 / 71.11	1.087	0.498	45.12	45.76	76.82 / 76.72	76.35 / 76.38	0.701	0.545
LNLN† 2024	34.31	38.06	73.34 / 72.05	73.75 / 72.19	1.059	0.525	45.42	46.17	76.30 / 78.19	77.77 / **79.95**	0.692	0.530
MASCF† 2025	29.80	32.19	70.52 / 68.55	67.99 / 65.77	1.078	0.510	44.69	45.26	72.24 / 74.46	68.17 / 71.38	0.722	0.490
TF-Mamba† 2025	33.35	36.21	73.94 / 73.06	73.82 / 73.04	1.055	0.541	45.66	46.64	77.34 / 77.61	77.18 / 77.43	0.673	0.578
RECAP (Ours)	**36.06**	**40.23**	**74.60 / 73.28**	**75.15 / 73.41**	**1.030**	**0.544**	**47.23**	**48.25**	**78.35 / 78.28**	**79.06** / 79.30	**0.659**	**0.599**

Table 1: Comparison of the average performance on the MOSI and MOSEI benchmarks under different missing rates (0.0-0.9). Models with † are reproduced under the same conditions. The best results are in bold, while the second-best are underlined.

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
MISA 2020	31.53	56.87	72.71	66.30	0.539	0.348
Self-MM 2021	32.28	56.75	72.81	68.43	0.508	0.376
MMIM 2021	31.81	52.76	69.86	66.21	0.544	0.339
TFR-Net 2021	26.52	52.89	68.13	58.70	0.661	0.169
CENET 2022	22.29	53.17	68.13	57.90	0.589	0.107
TETFN 2023	33.42	56.91	73.58	68.67	0.505	0.387
ALMT 2023	20.00	45.36	69.66	72.76	0.561	0.364
BI-Mamba 2024	31.90	54.95	70.79	69.26	0.529	0.345
LNLN 2024	34.64	57.14	72.73	**79.43**	0.514	0.397
MASCF 2025	29.33	53.36	70.67	69.96	0.507	0.402
TF-Mamba 2025	34.46	55.51	**74.68**	72.20	0.512	0.386
RECAP (Ours)	**37.02**	**59.56**	74.37	78.58	**0.499**	**0.417**

Table 2: Average performance comparison on CH-SIMS benchmark under various missing rates (0.0-0.9).

der varying degrees of modality incompleteness. Our experimental results are averages across five random seeds. RECAP consistently outperforms all competitors across nearly all metrics and datasets, demonstrating superior robustness and precision in modeling incomplete affective signals. On MOSI, RECAP achieves Acc-7 of 36.06% and Acc-5 of 40.23%, surpassing TF-Mamba by 2.71% and 4.02%, respectively. It also delivers the highest F1 score of 75.15% and the lowest MAE of 1.030, confirming its effectiveness in both categorical and regression settings. On MOSEI, RECAP attains the top performance with Acc-7 of 47.23%, F1 of 79.06%, and Corr of 0.599, improving over LNLN by 1.81%, 1.29%, and 0.069, respectively. On CH-SIMS, a dataset with fine-grained multilingual annotations, RECAP delivers a remarkably strong performance, reaching Acc-3 of 59.56% and F1 of 78.58%, outperforming TF-Mamba by 4.05% and 6.38%, respectively. Importantly, the consistent gains on high-resolution classification metrics indicate RECAP's ability to retain nuanced affective signals.

Ablation Study

We conduct ablation studies to assess the impact of each key component in RECAP. As shown in Table 3, removing any of these components leads to performance drops across both datasets, confirming their necessity. The absence of H-CAG results in the most pronounced degradation, with a point of 4.54 decline in F1 on SIMS and significantly reduced metrics on MOSI, indicating its crucial role in modeling temporal dependencies. FID also proves essential, as removing either shared or unique branches leads to noticeable declines. For instance, removing the unique branch causes a point of 1.4 drop in Acc-7 on MOSI and a point of 4.27 decline in F1 on SIMS, indicating the importance of modality-specific cues. MIR also plays a stabilizing role, as its removal results in a point of 1.25 drop in Acc-7 on MOSI and a point of 1.7 decline in F1 on SIMS, showing its value in adaptive fusion. Overall, these results demonstrate the effectiveness of our design in enhancing multimodal understanding.

Model	MOSI			SIMS		
	Acc-7	Acc-2	Corr	F1	MAE	Corr
RECAP (Ours)	35.07	**73.88**	**0.553**	**76.06**	**0.513**	**0.413**
w/o H-CAG	34.11	72.52	0.519	71.52	0.526	0.388
w/o FID	33.67	71.85	0.540	74.33	0.524	0.400
- w/o Shared	34.30	72.59	0.544	72.37	0.533	0.389
- w/o Unique	33.67	72.37	0.524	71.79	0.521	0.391
- w/o LV	34.11	72.18	0.534	72.14	0.515	0.392
- w/o LA	34.26	72.74	0.507	71.64	0.518	0.388
- w/o AV	**35.13**	71.47	0.517	74.62	**0.513**	0.400
w/o Reconstructor	34.40	71.40	0.518	74.13	0.514	0.398
w/o MIR	33.82	72.14	0.538	74.36	0.518	0.403

Table 3: Ablation study of different modules and strategies on MOSI and SIMS datasets. "L", "A", and "V" denote the language, acoustic, and visual modalities, respectively.

Results Analysis

Effect of Various Missing Rates. Figure 4 compares RECAP and LNLN under increasing levels of modality incompleteness ($r = 0, 0.2, 0.4, 0.6, 0.8$). As the missing rate increases, overall performance declines, with reductions in Acc-2, Acc-5, and Corr, alongside rising MAE. Subfigure (a) shows that RECAP consistently achieves higher Acc-2 and Corr, reflecting stronger robustness. Similarly, subfigure (b) further demonstrates RECAP's advantage in Acc-

5 and MAE across all missing rates, indicating more stable fine-grained prediction. These results confirm RECAP's resilience to missing data and its capacity to preserve fine-grained affective information even under severe degradation.

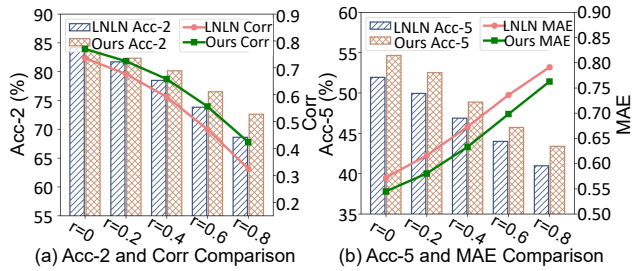

Figure 4: Model performance between LNLN (blue bars, pink lines) and our model (orange bars, green lines) with varying rates of missing modality data on MOSEI dataset.

Fine-Grained Analysis. We explore the discriminative capability of our model by visualizing the seven-class confusion matrices on the MOSI dataset with a 50% missing rate. As shown in Figure 5, RECAP (subfigure (b)) yields a more sharply diagonal and reduced off-diagonal dispersion matrix compared to MASCF (subfigure (a)), indicating stronger correspondence between predictions and ground truth. Notably, misclassifications of RECAP are primarily concentrated around adjacent sentiment levels, particularly at the extremes, suggesting that RECAP learns a more structured and continuous affective space that supports coherent and fine-grained prediction under multimodal input degradation.

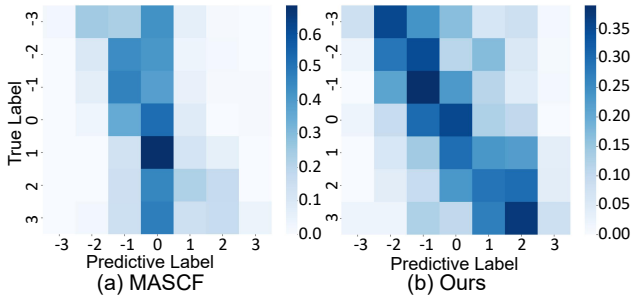

Figure 5: Seven-class confusion matrices of MASCF (left) and RECAP (right) on the MOSI dataset under a missing rate of 0.5. Labels -3 to 3 represent sentiment levels from strongly negative to strongly positive.

Visualization

Figure 6 provides a comprehensive visualization of the latent structure and generative behavior of RECAP under incomplete modality conditions on MOSI dataset. Subfigures (a) and (b) display the latent space learned by the FID module at a missing rate of 0.5. In subfigure (a), the similarity (shared) subspace shows tightly clustered language (orange) and visual (blue) features, reflecting consistent inter-modal semantics. In contrast, subfigure (b) reveals that the characteristic (modality-specific) subspace maintains clear separation, indicating successful disentanglement. These patterns confirm that FID could extract both commonality and specificity in a self-supervised manner even under incomplete conditions, facilitating more robust and structurally informed reconstruction. Subfigures (c) and (d) visualizes the distributions of generated (pink) and real (blue) modality features at missing rates of 0.5 and 0.9, respectively. At $r = 0.5$, we observe that the model tends to generate distributions that align closely with the ground-truth, exhibiting significant structural similarity and overlapping density regions. Even under extreme degradation $r = 0.9$, the generator still captures meaningful cues. The observed similarity suggests that the generator retains essential patterns, supporting plausible modality recovery under sparse conditions.

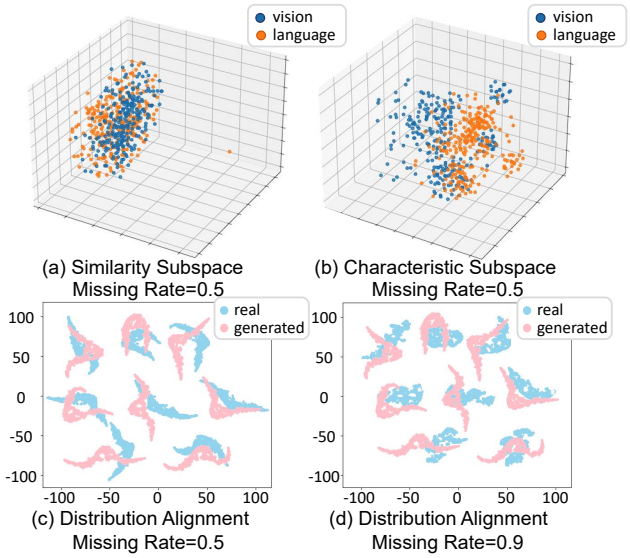

Figure 6: t-SNE visualization of learned latent spaces and generation quality. Subfigures (a) and (b) show similarity and characteristic subspaces learned by FID at a missing rate of 0.5. Orange and blue points represent language and visual modalities, respectively. Subfigures (c) and (d) illustrate the alignment between generated (pink) and real (blue) features under missing rates of 0.5 and 0.9, respectively.

Conclusion

This paper presents RECAP, a new framework tailored to MSA under real-world conditions where modalities are often incomplete or unreliable. By jointly modeling temporal causality and structural integrity, RECAP effectively reconstructs missing signals while preserving affective coherence. Our method also facilitates selective fusion through task-informed attention over modality cues. Both quantitative and qualitative results on multiple benchmark datasets demonstrate the effectiveness of RECAP in improving prediction accuracy and resilience to missing data, underscoring the importance of coherence-aware modeling in advancing MSA toward more practical and scalable applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62192781, 62576268, and 62137002, the Key Research and Development Project in Shaanxi Province No. 2023GXLH-024, the Project of China Knowledge Centre for Engineering Science and Technology, and the National Social Science Fund of China under Grant No. 23CWW006.

References

Achille, A.; and Soatto, S. 2018. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2897–2905.

Amjad, R. A.; and Geiger, B. C. 2019. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9): 2225–2239.

Chen, Q.; Tang, Y.; and Liu, H. 2025. Mamba-assisted modality subspace complementary fusion for multimodal sentiment analysis. *Pattern Recognition Letters*.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, 1779–1788. PMLR.

Deng, Z.; Ananthram, A.; and McKeown, K. 2025. Enhancing multimodal affective analysis with learned live comment features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16253–16261.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15180–15190.

Grau, M.; Lontke, A.; Jiang, X.; and Scheibenreif, L. 2023. Self supervised learning in remote sensing: Quantifying approaches effectiveness across downstream tasks. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, 518–521. IEEE.

Han, W.; Chen, H.; and Poria, S. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9180–9192.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

He, X.; Liang, H.; Peng, B.; Xie, W.; Khan, M. H.; Song, S.; and Yu, Z. 2025. MSAmba: Exploring Multimodal Sentiment Analysis with State Space Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1309–1317.

Jiang, Y.; Li, W.; Hossain, M. S.; Chen, M.; Alelaiwi, A.; and Al-Hammadi, M. 2020. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, 53: 209–221.

Lan, X.; Wu, F.; He, K.; Zhao, Q.; Hong, S.; and Feng, M. 2025. Gem: Empowering mllm for grounded ecg understanding with time series and images. *arXiv preprint arXiv:2503.06073*.

Li, J.; Savarese, S.; and Hoi, S. C. 2022. Masked unsupervised self-training for label-free image classification. *arXiv preprint arXiv:2206.02967*.

Li, M.; Yang, D.; and Zhang, L. 2023. Towards robust multimodal sentiment analysis under uncertain signal missing. *IEEE Signal Processing Letters*, 30: 1497–1501.

Li, X.; Cheng, X.; Miao, D.; Zhang, X.; and Li, Z. 2025. TF-Mamba: Text-enhanced Fusion Mamba with Missing Modalities for Robust Multimodal Sentiment Analysis. *arXiv e-prints*, arXiv–2505.

Liang, P. P.; Deng, Z.; Ma, M. Q.; Zou, J. Y.; Morency, L.-P.; and Salakhutdinov, R. 2023. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36: 32971–32998.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1): 309–325.

Sun, Y.; Liu, Z.; Sheng, Q. Z.; Chu, D.; Yu, J.; and Sun, H. 2024. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 110: 102454.

Tang, J.; Li, K.; Jin, X.; Cichocki, A.; Zhao, Q.; and Kong, W. 2021. CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on Natural Language Processing (volume 1: Long papers)*, 5301–5311.

Tao, C.; Li, J.; Zang, T.; and Gao, P. 2025. A Multi-Focus-Driven Multi-Branch Network for Robust Multimodal Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1547–1555.

Wang, D.; Guo, X.; Tian, Y.; Liu, J.; He, L.; and Luo, X. 2023a. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136: 109259.

Wang, D.; Liu, S.; Wang, Q.; Tian, Y.; He, L.; and Gao, X. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25: 4909–4921.

Wang, H.; Chen, Y.; Ma, C.; Avery, J.; Hull, L.; and Carneiro, G. 2023b. Multi-modal learning with missing

modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15878–15887.

Yang, Z.; Zhang, J.; Wang, G.; Kalra, M. K.; and Yan, P. 2024. Cardiovascular disease detection from multi-view chest x-rays with bi-mamba. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 134–144. Springer.

Yao, W.; Yin, K.; Cheung, W. K.; Liu, J.; and Qin, J. 2024. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 16416–16424.

Yi, G.; Fan, C.; Zhu, K.; Lv, Z.; Liang, S.; Wen, Z.; Pei, G.; Li, T.; and Tao, J. 2024. Vlp2msa: expanding vision-language pre-training to multimodal sentiment analysis. *Knowledge-Based Systems*, 283: 111136.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020. 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*, 3718–3727.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 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, 10790–10797.

Yuan, Z.; Li, W.; Xu, H.; and Yu, W. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM international conference on multimedia*, 4400–4407.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 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)*, 2236–2246.

Zeng, J.; Liu, T.; and Zhou, J. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1545–1554.

Zhang, H.; Wang, W.; and Yu, T. 2024. Towards robust multimodal sentiment analysis with incomplete data. *Advances in Neural Information Processing Systems*, 37: 55943–55974.

Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; and Yu, T. 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*, 756–767.

Zhang, Y.; Peng, C.; Wang, Q.; Song, D.; Li, K.; and Zhou, S. K. 2024. Unified multi-modal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging*, 44(1): 4–18.