CALIBRATE BEFORE DISENTANGLING: ROBUST MULTIMODAL SENTIMENT ANALYSIS UNDER MISSING OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust inference under corrupted observations is crucial for Multimodal Sentiment Analysis (MSA), as real-world textual, acoustic, and visual inputs are often incomplete or unreliable. Existing methods typically recover missing information through direct cross-modal compensation, but this paradigm leaves two key issues insufficiently addressed. First, unreliable observations can induce representation shifts that propagate through subsequent cross-modal interactions. Second, cross-modal evidence can facilitate the recovery of shared semantics but is inherently less suited to restoring modality-specific information. Motivated by these observations, we propose the **Ca**librated **Di**sentanglement **Net**work (CaDiNet), following a "calibrate-then-disentangle" paradigm. CaDiNet first performs reliability-aware calibration through a dual-granularity residual offset mechanism that leverages missingness information to capture global corruption patterns and instance-specific deviations before cross-modal interaction. It then disentangles each modality representation into shared and private components and refines them using component-specific evidence. Shared semantics are enhanced using reliable cross-modal information, whereas modality-specific information is refined using evidence retrieved from modality-homogeneous historical memory. Extensive experiments on four MSA benchmarks demonstrate that CaDiNet consistently improves robustness across diverse missing conditions.

1 INTRODUCTION

Multimodal Sentiment Analysis (MSA) aims to understand human language and behavior by jointly exploiting textual, visual, and acoustic cues (Zeng et al., 2024; Li & Tang, 2026; Wang et al., 2024; Kim et al., 2025), with broad applications in affective computing (Hazarika et al., 2020; Yang et al., 2022), human-agent interaction (Baruah & Banerjee, 2020), and social media analysis (Poria et al., 2019). Recent Transformer-based MSA models further leverage cross-modal complementarity through attention-based interaction (Tsai et al., 2019), modality-correlation modeling (Li et al., 2025), and representation enhancement (Wang et al., 2023; Ma et al., 2024; Quan et al., 2024). However, these methods typically assume complete and reliable multimodal observations, which may not hold in real-world scenarios due to sensor failures, privacy masking, and automatic speech recognition errors (Zhao et al., 2021; Yuan et al., 2024). Such disruptions lead to missing or corrupted inputs, making robust inference under incomplete observations a critical challenge for MSA.

Depending on the granularity of missingness, existing robust MSA methods can be broadly categorized into *modality-level* and *token-level* settings. Modality-level methods assume that one or more entire modalities are unavailable and typically recover missing information by reconstructing modality representations (Zhao et al., 2021; Zeng et al., 2023; Li et al., 2023; Huan et al., 2024) or learning missing-aware fused representations (Zhang et al., 2023; Guo et al., 2024; Li et al., 2024a;b). In contrast, token-level missingness involves unreliable observations scattered throughout multimodal sequences, making accurate recovery more challenging. Recent approaches improve robustness through noise imitation (Yuan et al., 2024), language-dominated correction (Zhang et al., 2024b), or text-guided reconstruction (Shi et al., 2025). Despite their effectiveness, these methods generally compensate for missing information by directly modeling cross-modal dependencies over

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

incomplete inputs, without calibrating missingness-induced representation shifts before cross-modal interaction or accounting for modality heterogeneity during compensation.

This recovery paradigm presents two key issues. First, corrupted observations exhibit representation shifts from their complete counterparts (Liu et al., 2025; Huang et al., 2025), with the magnitude governed by the missing ratio and the semantic direction determined by the instance context. Such missingness-induced representation shifts are misleading and cumulatively amplify across subsequent Transformer-based cross-modal interactions (see Appendix H). Second, multimodal representations are heterogeneous mixtures of modality-invariant (*i.e.*, shared) and modality-specific (*i.e.*, private) clues (Hazarika et al., 2020; Li et al., 2026) (see Section 4.3 for further discussion). Cross-modal evidence is mainly informative for recovering shared semantics, but it provides limited support for modality-specific details. Directly using observed modalities to reconstruct an entire missing representation therefore mixes shared information with private characteristics, thereby weakening modality-specific information and introducing inaccurate reconstructions.

To address these challenges, we propose the **Ca**librated **Di**sentanglement **Net**work (CaDiNet), instantiated through a "calibrate-then-disentangle" paradigm. Specifically, CaDiNet first calibrates missingness-induced representation shifts through a Dual-granularity Residual Offset (DRO) mechanism before Transformer-based cross-modal interaction, thereby mitigating error propagation. Furthermore, the calibrated representations are factorized into shared and private subspaces, allowing targeted refinement: shared semantics are enriched by dominant-modality evidence, whereas private characteristics are retrieved from modality-homogeneous historical memory rather than other modalities. Finally, the disentangled representations interact to produce robust predictions, enhancing both the semantic expressiveness and missing-modality robustness of the multimodal system.

The main contributions of this work are summarized as follows:

- We propose CaDiNet, a unified framework for MSA applicable to both token-level and modality-level missingness. It calibrates missingness-induced shifts before cross-modal interaction, then disentangles and refines shared and private representations with complementary evidence.
- We present an effective calibration mechanism, termed DRO, which consists of a global offset and an instance-specific offset. The former captures the dataset-level shift tendency under missingness, while the latter determines the instance-specific correction direction from reliable cross-modal evidence.
- CaDiNet disentangles the calibrated representations into shared and private subspaces, and refines them with dominant-modality evidence and retrieved private evidence, respectively, to enhance shared semantic consistency while preserving modality-specific characteristics.
- Extensive experiments on four MSA benchmarks validate the effectiveness of the "calibrate-then-disentangle" paradigm for reliable MSA. These findings provide useful insights for future research on incomplete MSA.

2 Related Work

2.1 Robust MSA under Incomplete Observations

Robust MSA methods mainly differ in the granularity of missingness. For modality-level missingness, one line explicitly reconstructs missing representations from available modalities through autoencoding, modality translation, or cross-modal attention (Zhao et al., 2021; Zeng et al., 2023; Huan et al., 2024; Li et al., 2023). Another line implicitly compensates for missing information at the fused-representation level using missing-aware tokens (Zhang et al., 2023), prompts (Guo et al., 2024), restoration (Li et al., 2024a), or distillation (Li et al., 2024b). These methods assume that missingness aligns with modality boundaries and are therefore less suited to token-level corruption, where missing entries are scattered within sequences and occur independently across modalities. For token-level missingness, recent methods employ adversarial reconstruction (Yuan et al., 2024), Transformer-based restoration (Yuan et al., 2021), language-guided correction (Zhang et al., 2024b), or text-guided cross-modal recovery (Shi et al., 2025). Despite these advances, most methods directly recover incomplete representations within an entangled multimodal space, without explicitly

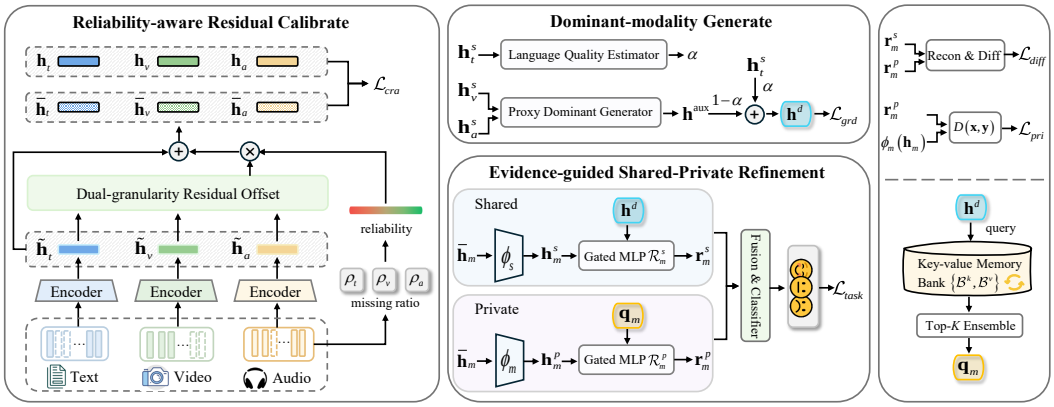


Figure 1: Overview of the proposed CaDiNet. Corrupted multimodal observations are first calibrated by the RRC module to mitigate missingness-induced representation shifts. The calibrated representations are then disentangled into shared and private subspaces and refined with dominant-modality and retrieved private evidence in the ESPR module.

calibrating missingness-induced representation shifts before cross-modal interaction, allowing corruption to propagate into subsequent fusion.

2.2 Disentangled Representation Learning

MISA (Hazarika et al., 2020) introduced a common paradigm of disentangling multimodal representations into shared and private subspaces. Subsequent works introduce more detailed constraints and task-specific refinement to better model shared consistency and private diversity (Sun et al., 2024; Wang et al., 2025). More recent studies further extend the binary shared-private decomposition toward multi-subspace modeling, including pairwise-shared and noise-related factors, enabling more detailed modeling of multimodal semantics (Meng et al., 2026; Yang & Li, 2026). However, most methods are developed for complete multimodal representations. A recent study, DERL (Li et al., 2026), extends disentangled learning to missing-modality scenarios through adaptive expert-based disentanglement and collaborative reconstruction. In contrast, our method first calibrates missingness-induced representation shifts before disentanglement, and then refines shared and private representations with distinct evidence sources.

3 Methodology

3.1 Workflow

The overall pipeline of CaDiNet is illustrated in Fig. 1. Incomplete multimodal observations are first encoded and calibrated by the Reliability-aware Residual Calibration (RRC) module using modality-specific global and instance-specific offsets. The calibrated representations are then disentangled into shared and private components, which are refined by the Evidence-guided Shared-Private Refinement (ESPR) module using dominant-modality and retrieved private evidence, respectively, before multimodal prediction.

3.2 Reliability-Aware Residual Calibration

Unlike prior methods that directly perform cross-modal interaction on missing embeddings, we calibrate missingness-induced representation shifts before subsequent Transformer-based interaction, preventing layer-wise error accumulation. We first formulate incomplete-input construction and modality encoding, and then introduce the DRO mechanism.

Incomplete-input Construction. Consistent with prior work (Huan et al., 2024; Zhang et al., 2024b), text is encoded using BERT (Devlin et al., 2019), while acoustic and visual features are ex-

tracted using Librosa (McFee et al., 2015) and OpenFace (Baltrusaitis et al., 2018), respectively. The pre-processed complete multimodal sample is written as $\mathbf{X} = \{\{\mathbf{x}_{m,j}\}_{j=1}^{L_m}\}_{m\in\mathcal{M}}$, where $\mathcal{M} \triangleq \{t, v, a\}$ denotes the set of modality indices corresponding to the textual, visual, and acoustic modalities, respectively, and L_m denotes the sequence length of modality m. To simulate incomplete observations, each sequence $\{\mathbf{x}_{m,j}\}_{j=1}^{L_m}$ is corrupted by a binary mask $\mathbf{o}_m \in \{0,1\}^{L_m}$, where $o_{m,j} = 1$ indicates that the j-th token is observed. The final corrupted input is constructed as:

$$\widetilde{\mathbf{x}}_{m,j} = o_{m,j}\mathbf{x}_{m,j} + (1 - o_{m,j})\mathbf{e}_m, \tag{1}$$

where $\mathbf{e}_m$ is the modality-specific missing placeholder, using `[UNK]` token for text and zero vectors for the other two continuous modalities. Masks are defined only over valid sequence positions, yielding the missing ratio $\rho_m = 1 - \frac{1}{L_m}\sum_{j=1}^{L_m} o_{m,j}$, where $\rho_m = 1$ indicates modality-level missing.

Modality encoding. We employ the BERT encoder for text and two-layer bidirectional recurrent encoders for the visual and acoustic inputs. $\mathcal{E}_t$ applies attention-mask mean pooling over BERT outputs, whereas $\mathcal{E}_v$ and $\mathcal{E}_a$ concatenate the bidirectional terminal states from both recurrent layers. A modality-specific projection $\mathcal{F}_m$ then maps every modality to a common d-dimensional space,

$$\widetilde{\mathbf{h}}_m = \mathcal{F}_m\Big(\mathcal{E}_m(\widetilde{\mathbf{X}}_m)\Big) \in \mathbb{R}^d. \tag{2}$$

The same encoding pipeline is applied to the complete inputs to obtain $\mathbf{h}_m$, which is used only for constructing the training objectives in Eqs. (5), (13), and (14).

Dual-granularity Residual Offset. We calibrate missingness-induced representation shifts using a DRO mechanism, consisting of a global offset and an instance-specific offset. The global offset is parameterized by a learnable modality-specific prompt $\boldsymbol{\pi}_m \in \mathbb{R}^d$ that captures the dataset-level shift tendency of modality m under missingness. The instance-specific offset instead adapts the correction to each sample by exploiting reliable evidence from the remaining modalities. Specifically, we construct a reliability-normalized context as

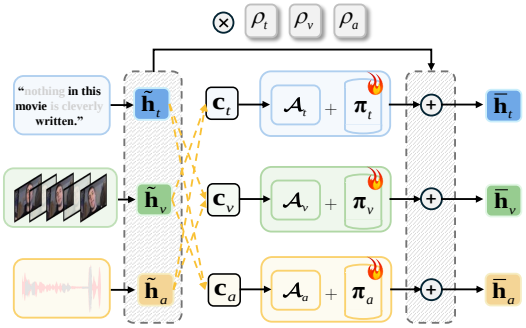


Figure 2: Illustration of the DRO mechanism.

$$\mathbf{c}_m = \frac{\sum_{n\neq m}\omega_n\widetilde{\mathbf{h}}_n}{\max\Big(\sum_{n\neq m}\omega_n,\ \epsilon\Big)}, \tag{3}$$

where ϵ is a small constant for numerical stability, and $\omega_m \triangleq p_m \cdot \mathcal{G}_m\Big([\widetilde{\mathbf{h}}_m; \rho_m]\Big) \in [0,1]$ denotes the reliability weight. Here, $p_m = 1 - \rho_m$ is the observed fraction, and $\mathcal{G}_m$ is a modality-specific gate implemented as a two-layer MLP with a sigmoid output. Their product downweights incomplete modalities and assigns zero weight to fully missing ones. The two offsets are then applied residually to calibrate the corrupted representation:

$$\overline{\mathbf{h}}_m = \widetilde{\mathbf{h}}_m + \rho_m\left(\boldsymbol{\pi}_m + \mathcal{A}_m(\mathbf{c}_m)\right), \tag{4}$$

where $\mathcal{A}_m$ maps the aggregated context to an instance-specific offset using a linear layer, a ReLU activation, and LayerNorm. The factor ρ_m adaptively scales the correction with the missing ratio and ensures $\overline{\mathbf{h}}_m = \widetilde{\mathbf{h}}_m$ when modality m is complete.

To further align the corrected representation $\overline{\mathbf{h}}_m$ with its complete-input counterpart $\mathbf{h}_m$, we introduce a complete-representation alignment loss:

$$\mathcal{L}_{cra} = \frac{1}{3N}\sum_{n=1}^{N}\sum_{m}\mathcal{D}\left(\overline{\mathbf{h}}_{m,n}, \mathbf{h}_{m,n}\right), \tag{5}$$

where N is the batch-size and $\mathcal{D}(\mathbf{x},\mathbf{y}) \triangleq \mathrm{SmoothL1}(\mathbf{x}-\mathbf{y}) + \eta[1 - \cos(\mathbf{x},\mathbf{y})]$ is a composite discrepancy metric. Here, $\mathrm{SmoothL1}(\cdot)$ measures magnitude discrepancy, while the cosine term captures directional misalignment, with $\eta = 0.1$ weighting the latter as a mild regularizer.

3.3 Shared-Private Disentanglement with Evidence-guided Refinement

Each calibrated representation is first decomposed into shared and private components using a shared encoder ϕ_s and a modality-specific encoder ϕ_m, respectively:

$$\mathbf{h}_m^s = \phi_s\left(\overline{\mathbf{h}}_m\right) = \sigma(\mathbf{W}_s\overline{\mathbf{h}}_m + \mathbf{b}_s), \qquad \mathbf{h}_m^p = \phi_m\left(\overline{\mathbf{h}}_m\right) = \sigma(\mathbf{W}_m\overline{\mathbf{h}}_m + \mathbf{b}_m). \tag{6}$$

This decomposition separates modality-invariant and modality-specific cues, allowing the two components to be refined with evidence tailored to their content. We first construct the dominant-modality evidence, which provides reliable cross-modal evidence for shared refinement and simultaneously serves as the semantic query for private evidence retrieval. We then use these complementary evidence sources to refine the disentangled representations for multimodal inference.

Dominant-modality Evidence Generation. Treating text as a fixed dominant modality can be unreliable under incomplete observations (Zhang et al., 2023; 2024b). We therefore generate adaptive dominant-modality evidence by complementing textual semantics with visual and acoustic information according to their reliability. Specifically, an auxiliary proxy is first generated from the reliability-scaled visual and acoustic representations together with their missing ratios:

$$\mathbf{h}^{\text{aux}} = \mathcal{P}([(1-\rho_v)\mathbf{h}_v^s; (1-\rho_a)\mathbf{h}_a^s; \rho_v; \rho_a]), \tag{7}$$

where $\mathcal{P}$ denotes the proxy generator. Meanwhile, a quality estimator $\mathcal{Q}$ evaluates the textual content, whose contribution is further bounded by its observed fraction $\alpha = \min(\mathcal{Q}(\mathbf{h}_t^s), 1-\rho_t)$. The dominant-modality evidence is then obtained as

$$\mathbf{h}^d = \alpha\mathbf{h}_t^s + (1-\alpha)\mathbf{h}^{\text{aux}}. \tag{8}$$

Accordingly, textual semantics dominate when text is reliable, whereas auxiliary modalities contribute more as textual reliability decreases. To ground $\mathbf{h}^d$ toward the complete textual representation $\mathbf{h}_t$, we further introduce an evidence-grounding loss $\mathcal{L}_{grd}$:

$$\mathcal{L}_{grd} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{D}\left(\mathbf{h}_n^d, \mathbf{h}_{t,n}\right). \tag{9}$$

Evidence-guided Shared-Private Refinement. For the shared components, $\mathbf{h}^d$ provides a common cross-modal semantic reference to refine each modality:

$$\mathbf{r}_m^s = \mathcal{R}_m^s\left(\mathbf{h}^d, \mathbf{h}_m^s\right), \tag{10}$$

where $\mathcal{R}_m^s$ is a modality-specific gated MLP that performs adaptive residual fusion.

Private components, however, are modality-specific and cannot be reliably recovered from other modalities. We maintain a key-value memory bank $\mathcal{B} = \{\mathcal{B}^k, \mathcal{B}^v\}$, where $\mathcal{B}^k \in \mathbb{R}^{N_{\text{mem}}\times d}$ stores semantic keys and $\mathcal{B}^v \in \mathbb{R}^{N_{\text{mem}}\times 3\times d}$ stores the aligned modality-specific private representations. As Figure 3 shows, using $\mathbf{h}^d$ as the query, we retrieve the top-K memory entries by cosine similarity and aggregate their values as

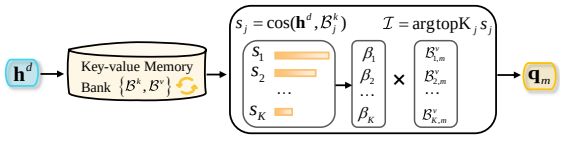


Figure 3: Illustration of private evidence retrieval.

$$\beta_j = \frac{\exp(s_j/\tau)}{\sum_{l\in\mathcal{I}}\exp(s_l/\tau)}, \qquad \mathbf{q}_m = \sum_{j\in\mathcal{I}}\beta_j\mathcal{B}_{j,m}^v, \tag{11}$$

where $s_j = \cos(\mathbf{h}^d, \mathcal{B}_j^k)$ and $\mathcal{I} = \arg\text{topK}_j\, s_j$. The retrieved evidence then refines the private representation:

$$\mathbf{r}_m^p = \mathcal{R}_m^p\left(\mathbf{q}_m, \mathbf{h}_m^p\right), \tag{12}$$

where $\mathcal{R}_m^p$ shares the architecture of $\mathcal{R}_m^s$ but uses separate parameters. During training, the memory bank $\mathcal{B}$ stores key-value pairs $\left\{\mathbf{h}^d, \{\phi_m(\mathbf{h}_m)\}_{m\in\mathcal{M}}\right\}$, whose values are private representations of the complete inputs. New pairs replace the oldest entries when the bank reaches its capacity N_{mem}. To avoid unstable retrieval from an insufficiently populated memory, private refinement is activated

only after a warm-up stage. During validation and testing, the memory bank $\mathcal{B}$ remains fixed and is used only for retrieval.

To preserve modality-specific semantics after evidence-guided refinement, we introduce a private consistency loss that aligns each refined private representation with its complete-input counterpart:

$$\mathcal{L}_{pri} = \frac{1}{3N} \sum_{n=1}^{N} \sum_{m} \mathcal{D}\left(\mathbf{r}_{m,n}^{p}, \phi_m\left(\mathbf{h}_{m,n}\right)\right). \tag{13}$$

We further regularize the disentangled representations with reconstruction and difference constraints to preserve information completeness and reduce redundancy between the shared and private components. The resulting disentanglement loss is defined as:

$$\mathcal{L}_{dis} = \frac{1}{3N} \sum_{n=1}^{N} \sum_{m} \left(\underbrace{\left\| \widehat{\mathbf{h}}_{m,n} - \mathbf{h}_{m,n} \right\|_2^2}_{\text{reconstruction constraint}} + \underbrace{\left[(\mathbf{r}_{m,n}^{s})^{\mathrm{T}} \mathbf{r}_{m,n}^{p} \right]^2}_{\text{difference constraint}} \right). \tag{14}$$

where $\widehat{\mathbf{h}}_{m,n} = \mathrm{Recon}\left(\mathbf{r}_{m,n}^{s} + \mathbf{r}_{m,n}^{p}\right)$ is the reconstructed representation.

Prediction and Joint Optimization. To capture inter-representation interactions, the six disentangled representations are concatenated into a sequence and fed into a lightweight Transformer encoder, followed by an MLP prediction head:

$$\hat{y} = \mathrm{MLP}\left(\mathrm{Transformer}\left([\mathbf{r}_t^p; \mathbf{r}_v^p; \mathbf{r}_a^p; \mathbf{r}_t^s; \mathbf{r}_v^s; \mathbf{r}_a^s]\right)\right). \tag{15}$$

Depending on the target task, the task-specific objective $\mathcal{L}_{task}$ is formulated as cross-entropy loss for classification or mean absolute error (MAE) for regression. The entire framework is trained end-to-end under the joint supervision of the task loss and multiple representation constraints:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_{cra}\mathcal{L}_{cra} + \lambda_{grd}\mathcal{L}_{grd} + \lambda_{pri}\mathcal{L}_{pri} + \lambda_{dis}\mathcal{L}_{dis}, \tag{16}$$

where $\lambda_{cra}, \lambda_{grd}, \lambda_{pri}$, and λ_{dis} are trade-off hyperparameters balancing the respective loss terms. A sensitivity analysis in Appendix D shows that the selected weights provide stable overall performance across different tasks.

4 Experiments

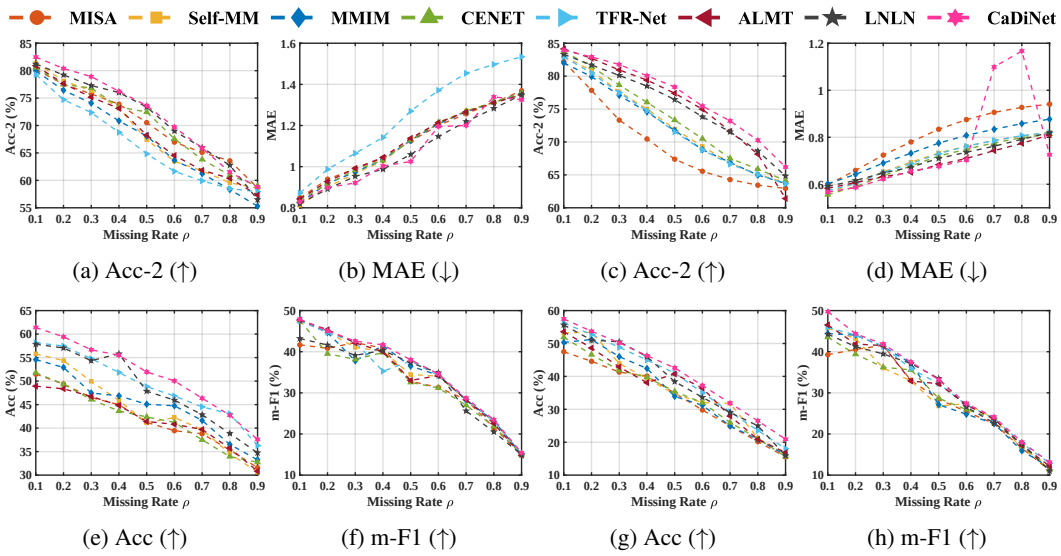
Datasets and Metrics. We evaluate CaDiNet on four benchmarks under diverse missing settings: CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018) for regression, and MELD-DA (Zhang et al., 2025; Poria et al., 2019) and MIntRec2.0 (Zhang et al., 2024a) for classification. For regression, we report binary accuracy (Acc-2), F1-score (F1), mean absolute error (MAE), and correlation (Corr), with Acc-2 and F1 computed under the negative versus positive setting; for classification, we report accuracy (Acc), macro-F1 (m-F1), macro-precision (m-P), and macro-recall (m-R). Higher values indicate better performance for all metrics, except MAE. Detailed descriptions, sample statistics, and feature extraction protocols for all datasets are deferred to Appendix B.1.

Baselines. We compare CaDiNet with representative baselines from three categories: self-supervised or information-theoretic multimodal representation learning methods, including Self-MM (Yu et al., 2021) and MMIM (Han et al., 2021); cross-modal interaction methods, represented by CENet (Wang et al., 2023); and robust methods for incomplete or noisy observations, including TFR-Net (Yuan et al., 2021), ALMT (Zhang et al., 2023), UniMF (Huan et al., 2024), and LNLN (Zhang et al., 2024b). Further details of these baselines are provided in Appendix B.2.

Implementation Details. Our framework is implemented using PyTorch and optimized via the AdamW optimizer with a batch size of 16. To ensure reproducibility and fair comparisons, all experiments are run across three fixed random seeds (1111, 1112, and 1113). Hyperparameter settings and experimental details are provided in Appendix C.

4.1 Robustness under Token-level Missingness

We first evaluate robustness under token-level missingness, where a common nominal missing rate ρ is applied to all modalities and varies from 0.1 to 0.9, yielding the modality-specific missing ratios



(a) Acc-2 (↑) (b) MAE (↓) (c) Acc-2 (↑) (d) MAE (↓)

(e) Acc (↑) (f) m-F1 (↑) (g) Acc (↑) (h) m-F1 (↑)

Figure 4: Performance on four benchmarks across varying token-level missing rates. (a)–(b) CMU-MOSI, (c)–(d) CMU-MOSEI, (e)–(f) MELD-DA, and (g)–(h) MIntRec2.0.

Table 1: Average performance on four benchmark datasets under token-level missingness.

Method	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
MISA	70.24	70.00	1.120	0.496	69.71	61.48	0.805	0.474	42.08	32.73	34.16	33.00	33.20	28.72	31.17	29.10
Self-MM	68.87	64.54	1.109	0.481	72.61	67.12	0.714	0.469	43.80	34.24	37.16	33.56	36.23	29.39	31.58	29.85
MMIM	67.55	64.79	1.114	0.477	72.19	67.09	0.758	0.463	44.78	34.34	37.24	33.72	35.22	29.91	32.44	30.40
CENET	70.23	66.85	1.117	0.473	73.47	69.05	0.702	0.509	42.07	32.52	35.65	32.66	34.47	29.07	31.78	29.25
TFR-Net	66.43	59.30	1.244	0.425	72.36	67.03	0.713	0.461	49.08	34.26	37.34	33.89	38.85	31.22	33.45	32.02
ALMT	68.79	71.19	1.120	0.468	75.64	76.19	**0.688**	0.451	41.93	34.24	36.39	34.00	36.23	30.57	33.51	31.29
LNLN	71.25	71.42	**1.079**	0.503	75.43	75.33	0.705	0.507	48.37	33.07	37.42	32.45	38.36	30.40	32.92	30.35
CaDiNet	**71.94**	**72.13**	1.082	**0.513**	**76.90**	**76.67**	0.756	**0.622**	**51.30**	**35.27**	**41.99**	**34.93**	**40.76**	**32.15**	**35.34**	**32.40**

ρ_m defined in Section 3.2. UniMF (Huan et al., 2024), which targets whole-modality absence, is therefore included only in the modality-level comparison (Section 4.2). Fig. 4 shows performance across missing rates, Table 1 reports the averages, and Appendix F provides the per-rate results.

As shown in Fig. 4, performance generally decreases as the missing rate increases, while CaDiNet remains consistently competitive across datasets. On CMU-MOSI, it achieves the best average Acc-2, F1, and Corr, with an MAE of 1.082 close to the best 1.079. On CMU-MOSEI, CaDiNet improves the second-best Acc-2 and F1 by 1.26 and 0.48 percentage points, respectively, and achieves the highest Corr of 0.622, although its MAE remains less competitive. For the more challenging classification datasets MELD-DA and MIntRec2.0, which contain more classes and stronger semantic ambiguity and class imbalance (Appendix B.1), CaDiNet achieves the best results across all four metrics. Overall, these results demonstrate stable robustness to token-level missingness, benefiting from representation calibration by RRC followed by evidence-guided refinement of the shared and private representations.

Table 2: Average performance on four benchmark datasets over seven input-availability cases.

Method	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
MISA	69.80	67.38	1.049	0.506	75.56	69.29	0.723	0.484	43.97	31.54	33.63	30.22	37.69	30.14	30.70	30.66
Self-MM	73.42	66.78	**1.005**	0.498	75.72	69.58	0.662	0.492	47.40	32.59	36.47	31.72	39.51	32.82	34.01	33.81
MMIM	70.50	66.24	1.013	0.492	74.66	70.29	0.710	0.488	47.40	30.64	34.28	30.66	39.03	33.11	32.92	33.70
CENET	72.19	65.57	1.022	0.498	75.17	70.86	0.662	0.521	45.21	30.58	35.69	31.15	40.82	33.05	34.74	33.93
TFR-Net	72.37	65.83	1.124	0.495	75.14	69.43	0.679	0.472	50.26	31.13	35.27	31.58	40.58	32.09	32.26	33.04
ALMT	72.58	**77.19**	1.076	0.493	64.22	72.79	0.687	0.411	47.30	31.37	37.83	31.26	39.44	33.31	33.61	33.67
UniMF	70.56	69.96	1.100	0.490	74.19	71.77	0.695	0.455	46.12	31.22	34.56	31.87	38.48	31.38	32.03	31.73
LNLN	70.77	73.81	1.044	0.474	75.04	**81.38**	0.710	0.469	50.06	31.61	36.08	31.90	42.18	33.41	33.74	33.16
CaDiNet	**74.72**	74.30	1.009	**0.541**	**77.39**	75.69	**0.651**	**0.542**	**51.70**	**34.15**	**39.49**	**34.58**	**43.69**	**35.24**	**36.78**	**36.02**

Table 3: Ablation study on model design and training objectives on four datasets with $\rho = 0.5$.

Variant	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
Full CaDiNet	**73.63**	**73.77**	**1.025**	**0.565**	**78.35**	**78.06**	0.676	**0.631**	**51.95**	38.03	42.47	36.77	**42.60**	**33.18**	34.73	**33.39**
Ablation of RRC																
w/o RRC	70.58	70.74	1.168	0.488	75.62	75.38	0.702	0.548	49.65	31.09	**49.27**	31.54	40.73	33.12	**36.31**	32.82
w/o Global Offset	71.49	71.66	1.141	0.476	77.35	77.42	0.679	0.578	50.85	37.86	42.57	36.64	41.37	32.22	36.27	32.95
w/o Instance-specific Offset	71.19	71.36	1.131	0.489	76.31	75.69	0.690	0.568	50.60	30.44	41.91	30.73	41.91	32.05	33.47	32.65
w/o Reliability Weighting	72.26	72.40	1.130	0.489	77.96	77.31	0.682	0.588	51.50	**38.53**	42.72	36.98	41.76	32.63	34.42	32.79
Ablation of Evidence-guided Refinement																
w/o Disentanglement	70.88	71.06	1.126	0.504	76.33	75.15	0.696	0.559	50.30	31.89	39.68	31.37	40.88	31.79	34.00	31.68
w/o Shared Refinement	72.10	72.26	1.121	0.510	77.16	77.37	0.668	0.595	50.45	33.59	43.87	34.01	41.91	32.05	33.47	32.65
w/o Private Refinement	71.95	72.11	1.147	0.477	76.75	75.56	0.680	0.585	50.65	36.98	40.68	36.82	41.47	30.81	33.34	31.58
Text-only Dominant Evidence	73.17	73.32	1.096	0.520	77.88	77.14	0.677	0.583	51.45	35.37	38.97	34.86	41.37	32.22	36.27	32.95
Ablation of Training Objectives																
w/o $\mathcal{L}_{cra}$	72.87	73.03	1.144	0.487	77.38	76.66	0.679	0.578	51.35	33.26	38.24	31.95	40.53	32.23	35.66	31.91
w/o $\mathcal{L}_{grd}$	71.04	71.10	1.169	0.502	76.94	77.22	**0.667**	0.599	51.15	35.38	37.99	34.76	41.42	32.17	35.20	32.36
w/o $\mathcal{L}_{pri}$	72.56	72.72	1.146	0.492	76.53	76.57	0.680	0.571	50.70	33.94	37.47	33.71	41.17	31.91	35.16	31.51
w/o $\mathcal{L}_{dis}$	71.34	71.28	1.154	0.443	77.05	77.16	0.669	0.588	50.95	38.43	42.29	**37.74**	40.63	30.69	34.43	31.36

4.2 Generalization to Modality-level Missingness

We further evaluate CaDiNet under modality-level missingness, where one or two complete modalities are removed during inference. Together with the complete-input setting, this evaluation covers seven input-availability cases: LAV, VA, LA, LV, V, A, and L. Table 2 summarizes the average performance over these cases, with the detailed results reported in Appendix G.

As shown in Table 2, CaDiNet generalizes effectively from token-level corruption to modality-level missingness. On CMU-MOSI, it achieves the best average Acc-2 and Corr, the second-best F1, and an MAE only 0.004 above the best result. On CMU-MOSEI, it ranks first in Acc-2, MAE, and Corr and second in F1, with the strong Corr indicating that sentiment relationships are well preserved under complete modality absence. On MELD-DA and MIntRec2.0, CaDiNet achieves the best results across all four metrics. Compared with the strongest competing results, it improves Acc and m-F1 by 1.44 and 1.56 percentage points on MELD-DA, and by 1.51 and 1.83 percentage points on MIntRec2.0, respectively. These results demonstrate robust generalization to modality-level missingness, where complementary cross-modal information supports shared refinement and retrieved historical evidence compensates for missing modality-specific information.

4.3 Ablation Studies

To examine the contribution of different designs in CaDiNet, we conduct ablation experiments at a missing ratio of $\rho = 0.5$. Table 3 evaluates the main components of RRC, evidence-guided representation refinement, and the training objectives on four datasets.

Ablation of RRC. Without RRC, the overall performance consistently drops across all four datasets, demonstrating the importance of calibrating missingness-induced representation shifts before subsequent cross-modal interaction. Further ablations show that both the global and instance-specific offsets contribute to the final performance. The former captures modality-level shift patterns, while the latter adapts the calibration to individual samples. Performance also generally decreases when reliability weighting is disabled, indicating that cross-modal evidence should be incorporated according to its reliability rather than treated equally.

Ablation of Evidence-guided Refinement. The disentanglement component is important for subsequent refinement, as its removal leads to clear performance degradation. Disabling either the shared or private refinement branch also generally reduces performance, showing that the two branches provide complementary information for recovering incomplete representations. Moreover, using the uncorrected textual representation as dominant-modality evidence performs worse than the full model, suggesting that auxiliary modalities provide useful information for correcting unreliable textual semantics under incomplete observations.

Ablation of Training Objectives. For the training objectives, excluding any auxiliary loss generally weakens the overall performance, despite occasional improvements on individual metrics. These objectives constrain different stages and representation spaces in CaDiNet, thereby provid-

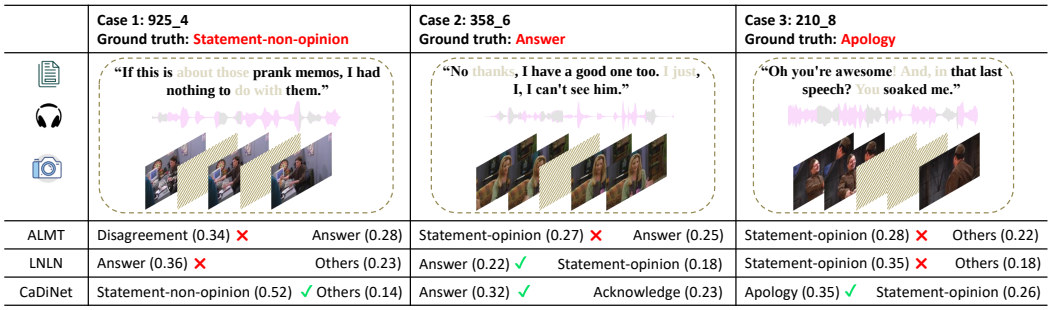


Figure 5: Qualitative comparison on MELD-DA under incomplete observations. Darkened regions denote missing observations and top-2 predictions and probabilities are reported.

ing complementary supervision during training. Their joint optimization is therefore beneficial for learning reliable multimodal representations under incomplete observations.

Table 4: Ablation of evidence assignment for shared and private refinement on CMU-MOSI and CMU-MOSEI.

Shared Evidence	Private Evidence	CMU-MOSI				CMU-MOSEI			
		Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)
$\mathbf{h}^d$	$\mathbf{q}_m$	**73.63**	**73.77**	**1.025**	**0.565**	**78.35**	**78.06**	**0.676**	**0.631**
$\mathbf{q}_m$	$\mathbf{q}_m$	68.75	68.82	1.228	0.458	75.62	75.38	0.702	0.548
$\mathbf{h}^d$	$\mathbf{h}^d$	69.66	69.83	1.200	0.451	75.92	76.11	0.683	0.573

Fine-grained Analysis of Evidence Assignment. We further examine whether shared and private refinement require different evidence sources. As shown in Table 4, assigning dominant-modality evidence $\mathbf{h}^d$ to the shared space and retrieved modality-specific evidence $\mathbf{q}_m$ to the private space achieves better overall performance than using either evidence source for both subspaces. This indicates that the two evidence sources are not interchangeable, and that effective refinement depends on matching the evidence source to the semantic role of each representation space.

5 Case Study

Figure 5 presents several MELD-DA examples under incomplete observations. These utterances are difficult to classify from textual form alone because their communicative functions can be close to other dialogue-act categories. For example, "No thanks" can be confused with disagreement, while an implicit apology may resemble a statement-opinion without clear lexical cues. The baseline models are more easily affected by such ambiguous textual semantics, whereas CaDiNet assigns the highest probability to the correct class across these cases. This suggests that the proposed calibration and evidence-guided refinement better preserve complementary multimodal cues under missing inputs. Additional case studies on MOSI, MOSEI, and MIntRec2.0 are provided in Appendix E.

6 Conclusion and Future Work

In this work, we propose CaDiNet for robust multimodal sentiment analysis under missing inputs. By combining reliability-aware residual calibration with evidence-guided disentangled representation refinement, CaDiNet reduces missingness-induced representation shifts and recovers useful modality-invariant and modality-specific information from reliable evidence. Extensive experiments on four benchmark datasets show that CaDiNet performs robustly under different token-level missing rates and generalizes well to complete modality absence. These results indicate that reliability-aware correction and differentiated recovery of modality-invariant and modality-specific information are important for robust multimodal learning. In future work, we will further study more complex corruption patterns and improve the generalization of CaDiNet to unseen missing conditions and other multimodal scenarios.

AI USE STATEMENT

In this work, we used generative AI tools (specifically, ChatGPT) to assist with English language polishing, grammatical corrections, and improving the readability of the manuscript. We also occasionally used AI-assisted tools for debugging experimental scripts and formatting LaTeX tables. We have not used generative AI tools for formulating core research ideas, deriving mathematical formulations, designing model architectures, generating raw experimental results, or producing scientific claims.

All AI-assisted outputs were carefully reviewed, verified, and edited by the human authors to ensure scientific accuracy, original tone, and factual correctness. All generated code was thoroughly tested and validated. We take full responsibility for the entirety of this work, including all text, claims, analyses, and results presented in this paper.

REFERENCES

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018.

Murchana Baruah and Bonny Banerjee. A multimodal predictive agent model for human interaction generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *ACM International Conference on Multimedia (ACM MM)*, 2020.

Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. UniMF: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Transactions on Multimedia*, 26, 2024.

Wuliang Huang, Yiqiang Chen, Xinlong Jiang, Chenlong Gao, Teng Zhang, Qian Chen, and Yifan Wang. Mitigating pervasive modality absence through multimodal generalization and refinement. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Tae-Young Kim, Jufeng Yang, and Eunil Park. MSDLF-K: A multimodal feature learning approach for sentiment analysis in korean incorporating text and speech. *IEEE Transactions on Multimedia*, 27, 2025.

Mingcheng Li, Dingkang Yang, and Lihua Zhang. Towards robust multimodal sentiment analysis under uncertain signal missing. *IEEE Signal Processing Letters*, 30, 2023.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024a.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *International Journal of Computer Vision*, 134(3), 2026.

Xiang Li, Xiaoming Zhang, Dezhuang Miao, Xianfu Cheng, Dawei Li, Honggui Han, and Zhoujun Li. Mixture of disentangled experts with missing modalities for robust multimodal sentiment analysis. *Information Fusion*, 135, November 2026.

Yangmin Li, Ruiqi Zhu, and Wengen Li. CorMulT: A semi-supervised modality correlation-aware multimodal transformer for sentiment analysis. *IEEE Transactions on Affective Computing*, 16 (3), July–September 2025.

Rui Liu, Haolin Zuo, Zheng Lian, Hongyu Yuan, and Qi Fan. Hardness-aware dynamic curriculum learning for robust multimodal emotion recognition with missing modalities. In *ACM International Conference on Multimedia (ACM MM)*, 2025.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26, 2024.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. M-SENA: An integrated platform for multimodal sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, 2022.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. Librosa: Audio and music signal analysis in python. In *Python in Science Conference (SciPy)*, 2015.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Weize Quan, Yunfei Feng, Ming Zhou, Yunzhen Zhao, Tong Wang, and Dong-Ming Yan. TCAN: Text-oriented cross attention network for multimodal sentiment analysis. *arXiv preprint arXiv:2404.04545*, 2024.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

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

Haoqin Sun, Shiwan Zhao, Xuechen Wang, Wenjia Zeng, Yong Chen, and Yong Qin. Fine-grained disentangled representation learning for multimodal emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

D. Wang, C. Tian, X. Liang, L. Zhao, L. He, and Q. Wang. Dual-perspective fusion network for aspect-based multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26, 2024.

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, 2023.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: Disentangled-language-focused multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In *ACM International Conference on Multimedia (ACM MM)*, 2022.

Zhilu Yang and Mingcheng Li. Factorize, reconstruct, enhance: A unified framework for multimodal sentiment analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *ACM International Conference on Multimedia (ACM MM)*, 2021.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26, 2024.

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

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia*, 25, 2023.

Ying Zeng, Wenjun Yan, Sijie Mai, and Haifeng Hu. Disentanglement translation network for multimodal sentiment analysis. *Information Fusion*, 102, February 2024.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *International Conference on Learning Representations (ICLR)*, 2024a.

Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. Multimodal classification and out-of-distribution detection for multimodal intent understanding. *IEEE Transactions on Multimedia*, 27, 2025.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

APPENDIX CONTENTS

- A. **Open resources**
- B. **Dataset and baseline details**
  - B.1. Dataset details
  - B.2. Baseline details
- C. **CaDiNet implementation details**
- D. **Sensitivity to loss weights**
- E. **Additional case studies**
- F. **Detailed results under token-level missingness**
- G. **Detailed results under modality-level missingness**
- H. **Mechanistic analysis of missingness-induced representation shifts**
  - H.1. Experimental setup
  - H.2. Layer-wise analysis of representation shifts
  - H.3. Effect on downstream prediction

A OPEN RESOURCES

To facilitate reproducibility and future research, we will publicly release the source code, model checkpoints, training and evaluation scripts, and hyperparameter configurations upon publication. The released materials will include detailed instructions for reproducing the main experimental results. To preserve the anonymity of the double-blind review process, the repository link is not provided in the current submission.

B DATASET AND BASELINE DETAILS

B.1 DATASET DETAILS

In this section, we provide detailed descriptions and statistics for the four datasets evaluated in our experiments:

- **CMU-MOSI** (Zadeh et al., 2016) is a widely utilized video benchmark for utterance-level multimodal sentiment intensity analysis. It consists of 2,199 video segments extracted from 93 YouTube monologue reviews, annotated with continuous sentiment intensity scores ranging from -3 (strongly negative) to $+3$ (strongly positive).
- **CMU-MOSEI** (Zadeh et al., 2018) scales up the diversity and complexity of MOSI in a naturalistic setting, containing 22,856 video utterances from 1,000 distinct speakers spanning 250 diverse topics. Each utterance is likewise labeled with a sentiment intensity score in $[-3, +3]$. Both MOSI and MOSEI serve to evaluate the robustness of our framework under continuous regression targets with incomplete multimodal inputs.
- **MELD-DA** (Zhang et al., 2025) is derived from the conversational emotion benchmark MELD (Poria et al., 2019), adapted from the TV show *Friends*. It contains over 13,000 utterances across 1,433 multiparty dialogues, annotated with dialogue act labels alongside synchronized textual, acoustic, and visual tracks.
- **MIntRec2.0** (Zhang et al., 2024a) is a large-scale, challenging benchmark designed for multimodal intent recognition in conversational contexts. In our experiments, we utilize its in-scope subset, which encompasses 9,304 utterance-level samples annotated with 30 fine-grained intent categories.

Label space and distribution. The label distributions of CMU-MOSI and CMU-MOSEI have been analyzed in prior work (Zhang et al., 2024b). Here, we focus on MELD-DA and MIntRec2.0, which present more challenging classification settings due to their larger label spaces, stronger semantic ambiguity, and class imbalance. As illustrated in Fig. 6, MELD-DA contains 12 dialogue-act classes, while MIntRec2.0 contains 30 intent classes. Many of these classes are semantically related and differ in subtle communicative functions, resulting in ambiguous class boundaries. For example, MELD-DA distinguishes between *statement-opinion* and *statement-non-opinion*, as well as closely related dialogue acts such as *agreement*, *acknowledgement*, and *backchannel*. Similarly, MIntRec2.0 contains fine-grained intents such as *inform* and *explain*, *agree* and *acknowledge*, and *greet* and *introduce*. Such semantic ambiguity makes distinguishing these classes more dependent on informative multimodal cues and therefore more challenging under incomplete observations.

Both datasets also exhibit pronounced class imbalance. In the MIntRec2.0 training set, the largest class contains 603 samples, whereas the smallest classes contain only 63, yielding an imbalance ratio of 9.57. Moreover, the seven classes with fewer than 100 samples account for only 8.21% of the training set. The imbalance is more pronounced in MELD-DA, where the largest class, *statement-non-opinion*, constitutes 29.98% of the training samples and is approximately 29 times larger than the smallest class, *acknowledge*. Its three largest classes jointly account for 63.29% of the training set, while the three smallest classes account for only 4.53%. In addition, MIntRec2.0 exhibits slight distribution shifts across data splits, whereas MELD-DA maintains nearly consistent class proportions. These characteristics make the two datasets particularly challenging for fine-grained classification under incomplete observations and also motivate the use of macro-averaged F1, precision, and recall alongside overall accuracy for a more comprehensive evaluation.

B.2 Baseline Details

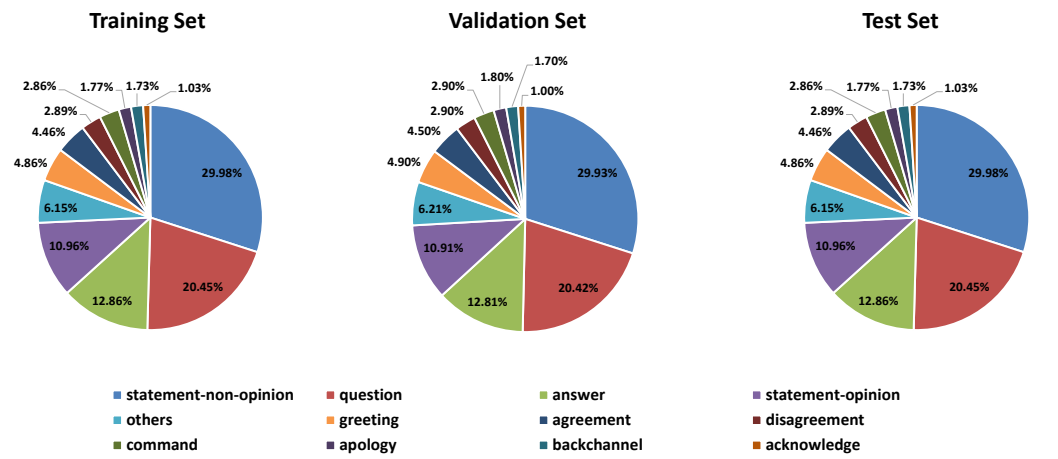
We provide brief descriptions of the compared baselines as follows.

Self-supervised or information-theoretic representation learning methods. Self-MM (Yu et al., 2021) adopts a self-supervised multi-task learning framework that generates unimodal pseudo-labels from multimodal supervision and jointly optimizes unimodal and multimodal tasks, encouraging the model to capture both cross-modal consistency and modality-specific characteristics. **MMIM** (Han et al., 2021) introduces hierarchical mutual information maximization to preserve task-relevant information during multimodal fusion, maximizing mutual information both across unimodal representations and between unimodal representations and the fused representation.

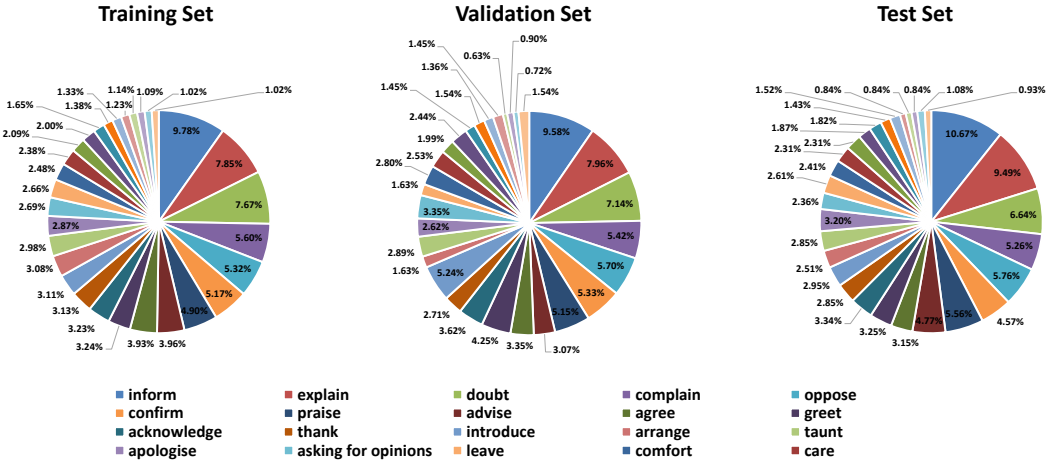
Cross-modal interaction methods. CENet (Wang et al., 2023) treats language as the dominant modality and introduces a cross-modal enhancement module into a pre-trained language model, where long-range visual and acoustic emotional cues are incorporated to enhance word-level textual representations.

Robust methods for incomplete or noisy observations. TFR-Net (Yuan et al., 2021) employs intra- and inter-modal attention to extract representations from incomplete multimodal sequences and reconstructs missing features under the supervision of their complete counterparts. **ALMT** (Zhang et al., 2023) introduces adaptive hyper-modality learning, where multi-scale language features guide visual and acoustic representations to suppress sentiment-irrelevant and conflicting information before multimodal fusion. **UniMF** (Huan et al., 2024) provides a unified framework for missing and unaligned multimodal sequences, using a generation Transformer to translate missing modalities from available ones and an understanding Transformer to integrate multimodal information for prediction. **LNLN** (Zhang et al., 2024b) regards language as the dominant modality and improves its reliability under incomplete observations through dominant-modality correction, followed by dominant-modality-based multimodal learning for robust prediction.

Overall, these baselines represent diverse strategies for multimodal representation learning, cross-modal interaction, and robustness to incomplete or noisy observations, providing a comprehensive basis for comparison with CaDiNet.



(a) Class distributions of MELD-DA.



(b) Class distributions of MIntRec2.0.

Figure 6: Class distributions across the training, validation, and test sets of MELD-DA and MIntRec2.0, illustrating their fine-grained, semantically related classes and pronounced class imbalance.

Table 5: Hyperparameter settings of CaDiNet on different datasets.

Hyperparameter	**MOSI**	**MOSEI**	**MIntRec2.0**	**MELD-DA**
Hidden Dimension d	128	128	128	128
Dropout Rate	0.2	0.2	0.2	0.2
Initial Learning Rate	2.5×10^{-5}	2.5×10^{-5}	1.5×10^{-5}	1.5×10^{-5}
Weight Decay	0.001	0.001	0.005	0.003
Maximum Epochs	100	100	30	20
Private Memory Size	256	512	512	512
Memory Retrieval Top-k	5	10	10	10
Loss Weights $[\lambda_{cra}, \lambda_{grd}, \lambda_{pri}, \lambda_{dis}]$	$[0.3, 0.4, 0.4, 0.2]$	$[0.3, 0.4, 0.5, 0.2]$	$[0.3, 0.4, 0.5, 0.2]$	$[0.3, 0.4, 0.4, 0.2]$
Random Seeds	1111, 1112, 1113	1111, 1112, 1113	1111, 1112, 1113	1111, 1112, 1113

C CaDiNet Implementation Details

We implement CaDiNet using PyTorch. The textual modality is encoded by `bert-base-uncased`[1], while the acoustic and visual modalities are encoded by two-layer

[1] `https://huggingface.co/google-bert/bert-base-uncased`

bidirectional LSTMs. All modality features are projected into a common 128-dimensional space. The fusion module contains one Transformer layer with two attention heads. The dropout rate and batch size are set to 0.2 and 16, respectively.

The model is optimized using AdamW with cosine learning-rate decay and a maximum gradient norm of 5.0. We train the model using both complete and incomplete observations, where the complete inputs serve only as stop-gradient supervision for representation repair. During inference, only incomplete inputs are required. Unless otherwise specified, the missing rate is set to 0.5 for validation and testing. Each experiment is repeated with three random seeds, *i.e.,* 1111, 1112, and 1113, and the average performance is reported. Model selection is based on validation MAE for MOSI and MOSEI and validation accuracy for MIntRec2.0 and MELD-DA.

For baseline methods, the results of MISA, Self-MM, MMIM, CENET, TETFN, TFR-Net, and ALMT are taken directly from Zhang et al. (2024b), which benchmarks them under default settings using the MMSA framework (Mao et al., 2022)[2]. The results for LNLN are reproduced using its official open-source repository[3]. The main hyperparameter configurations are summarized in Table 5.

The loss weights were selected according to preliminary experiments on the validation set, using validation MAE for CMU-MOSI and CMU-MOSEI and validation accuracy for MELD-DA and MIntRec2.0. We first determined a reasonable search range based on the relative scales of the loss terms and then varied one coefficient at a time while keeping the remaining coefficients fixed. Additional sensitivity results are provided in Appendix D.

D Sensitivity to Loss Weights

We investigate the sensitivity of CaDiNet to the weights of its four auxiliary objectives. Starting from the default configuration, we vary one coefficient at a time while keeping the remaining coefficients fixed. All configurations are evaluated under token-level missingness with $\rho = 0.5$, using the same data splits and training settings as the main experiments. The reported results are averaged over three random seeds.

Table 6: Sensitivity of CaDiNet to different loss-weight configurations on CMU-MOSI and CMU-MOSEI. The weight vector follows the order $[\lambda_{cra}, \lambda_{grd}, \lambda_{pri}, \lambda_{dis}]$. ↑ and ↓ indicate that higher and lower values are better, respectively.

Setting	CMU-MOSI					CMU-MOSEI				
	Loss Weights	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	**Loss Weights**	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)
Default	$[0.3, 0.4, 0.4, 0.2]$	<u>73.63</u>	<u>73.77</u>	**1.025**	<u>0.565</u>	$[0.3, 0.4, 0.5, 0.2]$	<u>78.35</u>	78.06	0.676	**0.631**
λ_{cra}: low	$[0.1, 0.4, 0.4, 0.2]$	73.48	73.63	1.078	0.548	$[0.1, 0.4, 0.5, 0.2]$	78.29	77.97	0.671	0.586
λ_{cra}: high	$[0.5, 0.4, 0.4, 0.2]$	73.02	73.15	1.135	0.539	$[0.5, 0.4, 0.5, 0.2]$	78.34	77.95	0.678	0.586
λ_{grd}: low	$[0.3, 0.2, 0.4, 0.2]$	73.17	73.30	1.148	0.492	$[0.3, 0.2, 0.5, 0.2]$	**78.67**	**78.26**	0.675	0.588
λ_{grd}: high	$[0.3, 0.6, 0.4, 0.2]$	**73.93**	**74.05**	1.111	0.549	$[0.3, 0.6, 0.5, 0.2]$	78.12	77.82	0.674	0.581
λ_{pri}: low	$[0.3, 0.4, 0.2, 0.2]$	73.32	73.48	1.091	0.557	$[0.3, 0.4, 0.3, 0.2]$	78.34	<u>78.21</u>	<u>0.663</u>	0.589
λ_{pri}: high	$[0.3, 0.4, 0.6, 0.2]$	73.48	73.63	<u>1.062</u>	**0.575**	$[0.3, 0.4, 0.7, 0.2]$	78.12	77.60	0.672	0.585
λ_{dis}: low	$[0.3, 0.4, 0.4, 0.1]$	73.02	73.02	1.143	0.562	$[0.3, 0.4, 0.5, 0.1]$	78.21	77.89	**0.654**	<u>0.613</u>
λ_{dis}: high	$[0.3, 0.4, 0.4, 0.4]$	73.17	73.33	1.099	0.545	$[0.3, 0.4, 0.5, 0.4]$	78.15	77.83	0.668	0.605

Tables 6 and 7 show that the selected configurations provide a balanced trade-off rather than uniformly optimizing every test metric. On CMU-MOSI, the default setting achieves the lowest MAE and remains close to the best Acc-2, F1, and Corr results. On CMU-MOSEI, it obtains the highest Corr while remaining competitive in Acc-2 and F1, although smaller λ_{grd} or λ_{dis} improves some individual metrics.

The classification results exhibit a similar pattern. The default setting achieves the highest accuracy on both MELD-DA and MIntRec2.0. Increasing λ_{dis} improves the macro-averaged metrics on MELD-DA, whereas a smaller λ_{grd} favors m-F1 and m-R on MIntRec2.0. In contrast, an overly

[2] https://github.com/thuiar/MMSA

[3] https://github.com/Haoyu-ha/LNLN

Table 7: Sensitivity of CaDiNet to different loss-weight configurations on MELD-DA and MIntRec2.0. The weight vector follows the order $[\lambda_{cra}, \lambda_{grd}, \lambda_{pri}, \lambda_{dis}]$. Higher values indicate better performance for all metrics.

Setting	MELD-DA					MIntRec2.0				
	Loss Weights	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Loss Weights	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
Default	$[0.3, 0.4, 0.4, 0.2]$	**51.95**	38.03	42.47	36.77	$[0.3, 0.4, 0.5, 0.2]$	**42.60**	33.18	34.73	33.39
λ_{cra}: low	$[0.1, 0.4, 0.4, 0.2]$	51.60	33.11	40.94	33.76	$[0.1, 0.4, 0.5, 0.2]$	41.27	32.63	33.83	32.93
λ_{cra}: high	$[0.5, 0.4, 0.4, 0.2]$	51.85	37.26	42.33	36.37	$[0.5, 0.4, 0.5, 0.2]$	42.30	33.73	33.57	34.55
λ_{grd}: low	$[0.3, 0.2, 0.4, 0.2]$	51.70	37.43	43.70	35.97	$[0.3, 0.2, 0.5, 0.2]$	42.35	**34.16**	35.91	**34.69**
λ_{grd}: high	$[0.3, 0.6, 0.4, 0.2]$	51.65	35.59	**45.93**	33.97	$[0.3, 0.6, 0.5, 0.2]$	42.55	31.48	37.18	31.77
λ_{pri}: low	$[0.3, 0.4, 0.2, 0.2]$	51.75	37.52	42.50	36.01	$[0.3, 0.4, 0.3, 0.2]$	42.35	33.06	33.84	33.97
λ_{pri}: high	$[0.3, 0.4, 0.6, 0.2]$	51.55	24.96	29.53	26.75	$[0.3, 0.4, 0.7, 0.2]$	42.40	32.62	35.03	34.15
λ_{dis}: low	$[0.3, 0.4, 0.4, 0.1]$	51.75	37.66	44.05	36.04	$[0.3, 0.4, 0.5, 0.1]$	42.30	33.73	33.57	34.55
λ_{dis}: high	$[0.3, 0.4, 0.4, 0.4]$	51.50	**38.53**	42.72	**36.98**	$[0.3, 0.4, 0.5, 0.4]$	42.30	33.63	**37.87**	34.10

large λ_{pri} causes a clear degradation in the macro metrics of MELD-DA, indicating that excessive private-consistency regularization can reduce class-wise balance on this long-tailed dataset. Overall, the results confirm that the preferred loss balance is task dependent. The selected weights offer stable overall performance, while the loss ablations in Table 3 further verify the complementary roles of the four auxiliary objectives.

E Additional Case Studies

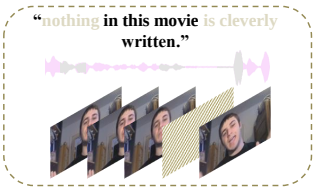
We provide additional qualitative comparisons on CMU-MOSI, CMU-MOSEI, and MIntRec2.0 to complement the MELD-DA examples reported in the main paper. Figures 7 and 8 present representative cases under incomplete observations. Missing observations are explicitly marked in the corresponding modalities. For each case, we report the incomplete multimodal inputs together with model predictions, including sentiment polarity with confidence for CMU-MOSI and CMU-MOSEI, and the top-2 predicted intent classes with probabilities for MIntRec2.0.

For CMU-MOSI and CMU-MOSEI, the selected cases contain complex sentiment expressions whose overall polarity cannot be reliably inferred from isolated lexical cues. For example, expressions such as "not all bad" and "wasn't annoying" involve negation, while "it's not as bad as I thought it would be" conveys a positive sentiment despite containing the locally negative expression "not as bad." Under incomplete observations, such subtle semantic compositions can be further obscured by missing textual tokens or nonverbal cues. As shown in Fig. 7, ALMT and LNLN frequently assign higher confidence to the incorrect polarity in these cases, whereas CaDiNet consistently recovers the correct predictions. This suggests that calibrating missingness-induced representation shifts before cross-modal interaction, followed by evidence-guided refinement, helps preserve complementary multimodal cues for resolving ambiguous sentiment semantics.

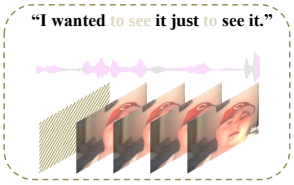
For MIntRec2.0, the difficulty mainly arises from fine-grained intent categories with semantically overlapping expressions. As shown in Fig. 8, the first example can be confused with *taunt* or *greet*, while the second exhibits ambiguity among *refuse*, *doubt*, and *explain*. In the third example, both LNLN and CaDiNet correctly identify *taunt*, but CaDiNet assigns a higher probability to the correct intent and produces *criticize* rather than *inform* as the second candidate, which is semantically closer to the sarcastic expression. These cases indicate that CaDiNet better preserves discriminative multimodal evidence for separating closely related intents when observations are incomplete.

F Detailed Results under Token-level Missingness

This appendix provides the complete per-rate results underlying the averaged comparison in Table 1 of the main paper. For clearer presentation, we divide the random token-level missing rates into three ranges: low missingness ($\rho = 0.1$–0.3), medium missingness ($\rho = 0.4$–0.6), and high missingness ($\rho = 0.7$–0.9).

	Case 1: f9O3YtZ2VfI$_$20 Ground truth: Positive	Case 2: tmZoasNr4rU$_$13 Ground truth: Negative	Case 3: jUzDDGyPkXU$_$19 Ground truth: Positive
			
ALMT	Negative (0.72) ✗ Positive (0.28)	Negative (0.59) ✗ Positive (0.41)	Negative (0.65) ✗ Positive (0.35)
LNLN	Negative (0.63) ✗ Positive (0.37)	Positive (0.53) ✗ Negative (0.47)	Negative (0.69) ✗ Positive (0.31)
CaDiNet	Positive (0.52) ✓ Negative (0.48)	Negative (0.52) ✓ Positive (0.48)	Positive (0.56) ✓ Negative (0.44)

(a) CMU-MOSI

	Case 1: 53609_11 Ground truth: Negative	Case 2: 245582_16 Ground truth: Positive	Case 3: d1CDP6sMuLA_9 Ground truth: Positive
			
ALMT	Negative (0.62) ✓ Positive (0.38)	Negative (0.77) ✗ Positive (0.23)	Negative (0.72) ✗ Positive (0.28)
LNLN	Positive (0.58) ✗ Negative (0.42)	Negative (0.73) ✗ Positive (0.27)	Negative (0.63) ✗ Positive (0.37)
CaDiNet	Negative (0.66) ✓ Positive (0.34)	Positive (0.67) ✓ Negative (0.33)	Positive (0.58) ✓ Negative (0.42)

(b) CMU-MOSEI

Figure 7: Qualitative comparisons on CMU-MOSI and CMU-MOSEI, showing the advantage of CaDiNet in handling incomplete observations with complex sentiment semantics.

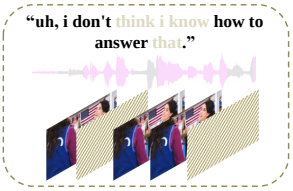
	Case 1: 332_5 Ground truth: Praise	Case 2: 435_3 Ground truth: Refuse	Case 3: 43_9 Ground truth: Taunt
			
ALMT	Taunt (0.34) ✗ Praise (0.28)	Doubt (0.27) ✗ Refuse (0.25)	Joke (0.28) ✗ Taunt (0.22)
LNLN	Greet (0.36) ✗ Taunt (0.23)	Explain (0.22) ✗ Oppose (0.18)	Taunt (0.35) ✓ Inform (0.18)
CaDiNet	Praise (0.52) ✓ Taunt (0.14)	Refuse (0.32) ✓ Acknowledge (0.23)	Taunt (0.35) ✓ Criticize (0.26)

Figure 8: Qualitative comparisons on MIntRec2.0, showing the advantage of CaDiNet in distinguishing semantically ambiguous intents under incomplete observations.

Low missingness. As shown in Table 8, all methods retain relatively strong performance when only a small portion of tokens is missing. CaDiNet nevertheless achieves consistently strong results across the four datasets, particularly on the main classification metrics. This indicates that the proposed calibration and refinement do not sacrifice performance when the available multimodal evidence remains largely complete.

Medium missingness. The differences between methods become more evident as the missing rate increases to 0.4–0.6 (Table 9). CaDiNet maintains strong performance on both sentiment regression and fine-grained classification, and achieves the best results on most metrics at $\rho = 0.5$ and $\rho =$

Table 8: Results under low token-level missingness ($\rho = 0.1$–0.3) show strong overall performance of CaDiNet.

Method	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
Random Missing Rate $\rho = 0.1$																
MISA	80.18	80.21	0.847	0.721	82.21	81.28	0.598	0.722	51.49	41.67	42.24	40.75	47.49	39.28	44.87	39.58
Self-MM	81.40	81.19	**0.812**	0.728	83.03	82.43	0.564	0.725	55.79	47.45	50.48	42.86	53.19	45.73	51.45	45.80
MMIM	79.98	79.83	0.825	0.718	82.00	81.57	0.602	0.696	54.57	47.82	50.35	44.23	50.34	44.42	48.99	44.62
CENET	80.59	80.55	0.837	0.719	83.75	83.42	**0.556**	0.739	51.80	47.59	49.11	47.26	51.92	43.54	46.94	43.34
TFR-Net	79.27	78.70	0.872	0.705	82.92	82.25	0.573	0.715	58.16	47.50	45.94	46.11	56.20	45.91	51.90	46.67
ALMT	80.90	81.15	0.843	0.703	**84.14**	**84.23**	0.583	0.718	48.93	47.70	46.89	45.37	53.57	46.52	**53.20**	46.90
LNLN	81.20	81.62	0.820	0.724	83.32	83.66	0.591	0.712	57.85	43.22	**50.80**	42.50	55.64	44.26	45.38	44.70
CaDiNet	**82.47**	**82.44**	0.829	**0.736**	83.93	83.79	0.567	**0.743**	**61.41**	**47.89**	50.33	**47.56**	**57.45**	**49.75**	51.54	**49.90**
Random Missing Rate $\rho = 0.2$																
MISA	77.54	77.58	0.939	0.654	77.84	75.56	0.659	0.674	49.38	40.84	41.07	40.74	44.59	40.55	41.20	42.48
Self-MM	78.15	77.76	0.901	0.660	80.84	79.76	0.604	0.678	54.36	44.76	48.21	43.65	53.46	43.14	44.14	43.52
MMIM	76.42	76.12	0.918	0.651	79.93	79.08	0.642	0.653	52.91	44.66	49.09	43.55	51.21	44.03	43.71	47.43
CENET	77.49	77.35	0.916	0.654	81.46	80.78	0.590	0.698	49.29	39.51	41.55	39.30	46.63	39.40	40.59	40.11
TFR-Net	74.70	73.57	0.987	0.622	80.47	79.29	0.604	0.672	57.47	44.71	48.41	43.60	52.73	44.08	45.54	**47.52**
ALMT	77.64	77.94	0.927	0.645	82.71	**82.82**	0.607	0.669	48.36	**45.36**	47.34	43.50	48.65	41.87	45.57	45.28
LNLN	79.22	79.53	**0.891**	**0.698**	81.70	81.95	0.616	0.677	57.05	41.62	49.95	41.10	51.04	41.26	42.18	41.60
CaDiNet	**80.39**	**80.40**	0.898	0.679	**82.92**	82.76	**0.586**	**0.714**	**59.41**	45.06	**50.25**	**43.95**	**53.71**	**44.38**	**45.87**	45.01
Random Missing Rate $\rho = 0.3$																
MISA	75.76	75.82	0.989	0.618	73.32	68.91	0.724	0.615	46.51	42.22	40.85	41.65	41.32	41.55	40.95	41.65
Self-MM	76.37	75.68	0.967	0.614	77.63	75.69	0.653	0.610	49.94	41.03	41.17	39.91	43.93	36.04	36.83	38.45
MMIM	74.08	73.47	0.974	0.612	77.08	75.46	0.690	0.597	47.47	37.79	38.34	37.35	46.00	41.24	43.59	41.60
CENET	76.83	76.56	0.983	0.605	78.65	77.34	0.636	0.640	46.09	38.15	43.13	37.33	41.85	36.24	41.46	36.47
TFR-Net	72.36	70.12	1.065	0.572	77.48	75.43	0.650	0.604	54.77	42.27	45.82	41.75	48.83	41.33	39.21	41.70
ALMT	75.15	75.51	0.992	0.596	80.94	81.15	0.632	0.598	46.74	42.17	40.58	41.70	42.88	41.50	42.07	41.55
LNLN	77.29	77.56	0.953	0.617	80.11	80.44	0.648	0.629	54.35	39.12	45.55	38.40	**50.52**	39.46	41.98	39.60
CaDiNet	**78.91**	**78.95**	**0.921**	**0.657**	**81.76**	**81.57**	**0.622**	**0.649**	**56.66**	**42.57**	**45.85**	**42.05**	50.22	**41.85**	**45.05**	**42.00**

Table 9: Results under medium token-level missingness ($\rho = 0.4$–0.6) show stable advantages of CaDiNet as corruption increases.

Method	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
Random Missing Rate $\rho = 0.4$																
MISA	73.88	73.88	1.041	0.585	70.46	64.02	0.780	0.561	44.98	40.88	42.85	40.20	40.09	33.02	34.18	34.83
Self-MM	73.17	71.74	1.027	0.579	75.02	72.01	0.694	0.554	45.89	40.15	42.19	39.75	39.37	32.89	34.03	33.55
MMIM	70.84	69.69	1.034	0.576	74.56	71.98	0.732	0.542	46.86	41.47	43.13	40.25	42.41	37.19	39.26	37.09
CENET	73.38	72.75	1.031	0.574	76.03	73.87	0.678	0.587	43.63	39.80	43.54	39.04	39.73	35.57	36.42	35.86
TFR-Net	68.75	64.71	1.142	0.537	74.74	71.67	0.688	0.548	51.83	35.27	36.43	35.06	44.97	35.85	35.79	37.23
ALMT	73.12	73.85	1.045	0.560	79.40	79.68	**0.651**	0.536	44.84	40.03	41.08	40.15	38.09	32.89	33.21	34.22
LNLN	76.01	76.31	**0.987**	0.594	78.49	78.98	0.673	0.592	**55.81**	40.32	42.30	39.50	45.99	37.14	40.98	37.04
CaDiNet	**76.22**	**77.15**	1.003	**0.599**	**80.06**	**79.81**	0.656	**0.603**	55.51	**41.66**	**47.44**	**40.55**	**46.29**	**37.49**	**41.31**	**37.39**
Random Missing Rate $\rho = 0.5$																
MISA	70.53	70.50	1.124	0.519	67.38	58.38	0.834	0.492	41.23	32.81	33.62	32.71	34.91	28.45	32.77	28.06
Self-MM	67.43	64.27	1.129	0.503	71.97	67.40	0.733	0.477	41.13	34.45	34.63	34.47	34.60	26.81	29.29	26.80
MMIM	68.09	66.15	1.128	0.501	71.75	67.70	0.775	0.470	45.08	36.49	38.84	36.00	33.93	27.14	29.39	27.05
CENET	72.46	71.10	1.130	0.496	73.33	69.80	0.720	0.515	42.39	32.56	33.05	32.65	35.41	28.68	31.85	28.19
TFR-Net	64.83	58.04	1.270	0.443	71.53	66.88	0.730	0.471	48.84	37.73	39.49	36.47	41.04	32.26	33.98	32.50
ALMT	68.24	69.74	1.138	0.485	77.40	77.73	0.683	0.461	41.40	33.02	33.95	32.97	40.74	32.21	33.88	32.52
LNLN	73.37	73.70	1.059	0.536	76.44	77.23	0.710	0.529	47.85	37.32	39.10	36.42	38.44	**33.48**	31.38	30.50
CaDiNet	**73.63**	**73.77**	**1.025**	**0.565**	**78.35**	**78.06**	**0.676**	**0.631**	**51.95**	**38.03**	**42.47**	**36.77**	**42.60**	33.18	**34.73**	**33.39**
Random Missing Rate $\rho = 0.6$																
MISA	66.97	66.94	1.200	0.441	65.55	54.64	0.875	0.415	39.45	31.32	33.32	32.63	29.79	25.44	27.44	25.04
Self-MM	63.47	58.94	1.209	0.425	69.33	63.01	0.762	0.401	42.26	34.06	37.94	32.77	32.06	27.35	28.03	27.31
MMIM	63.67	60.87	1.208	0.418	68.83	63.09	0.808	0.402	44.75	34.29	34.48	33.26	31.23	24.80	26.34	25.51
CENET	67.58	64.87	1.215	0.415	70.50	65.27	0.749	0.446	41.22	31.19	35.46	31.17	32.02	25.95	27.54	26.35
TFR-Net	61.64	52.44	1.371	0.363	68.80	62.51	0.762	0.397	46.83	34.24	39.26	33.21	35.04	27.09	30.34	27.59
ALMT	64.53	66.81	1.214	0.407	74.98	**75.44**	0.710	0.395	40.81	34.19	39.83	33.16	36.44	27.04	31.05	27.54
LNLN	69.00	69.19	**1.147**	0.458	73.82	75.03	0.736	0.471	45.95	**34.89**	31.60	29.30	33.64	26.56	31.08	26.80
CaDiNet	**69.71**	**69.90**	1.195	**0.485**	**75.46**	75.06	**0.703**	**0.527**	**50.05**	34.59	**40.42**	**33.56**	**37.24**	**27.39**	**31.35**	**27.89**

0.6. These results suggest that calibrating corrupted representations and refining shared and private information become increasingly important as reliable observations decrease.

High missingness. Table 10 shows that severe token-level corruption substantially reduces the performance of all methods. Despite this challenging setting, CaDiNet remains competitive on MOSI and performs particularly well on MOSEI, MELD-DA, and MIntRec2.0 across most metrics. The advantage is more evident on the two fine-grained classification datasets, where CaDiNet consistently achieves strong Acc and macro-level scores. This suggests that the proposed framework better preserves task-relevant information when only limited multimodal evidence remains.

Overall, the per-rate results show that CaDiNet remains effective across different levels of token-level missingness. Its advantage is especially clear under medium and high missing rates, supporting the effectiveness of the calibrate-then-disentangle paradigm for learning reliable multimodal representations from incomplete observations.

Table 10: Results under high token-level missingness ($\rho = 0.7$–0.9) show that CaDiNet remains strong under severe corruption.

Method	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
Random Missing Rate $\rho = 0.7$																
MISA	65.09	65.07	1.257	0.381	64.28	51.82	0.906	0.344	38.80	28.32	30.72	28.45	24.91	22.50	25.39	21.84
Self-MM	61.74	55.11	1.271	0.339	66.79	58.05	0.786	0.329	39.66	28.24	33.42	28.12	31.79	23.77	26.35	23.62
MMIM	61.23	57.15	1.267	0.342	66.89	58.90	0.834	0.341	41.63	28.29	33.64	28.17	24.93	22.53	26.48	21.59
CENET	63.82	53.79	1.269	0.335	67.50	59.88	0.776	0.384	37.52	27.00	29.82	27.31	25.99	23.72	26.40	23.91
TFR-Net	59.91	48.41	1.454	0.276	66.64	58.32	0.786	0.322	44.52	28.39	35.63	28.27	29.21	23.82	26.45	23.52
ALMT	61.84	65.30	1.266	0.336	71.62	72.24	**0.743**	0.315	39.72	28.34	33.72	**28.87**	27.93	23.67	26.30	23.57
LNLN	**65.95**	65.95	1.219	0.383	71.55	71.52	0.762	0.408	42.85	25.62	29.40	25.80	29.04	22.66	25.28	22.80
CaDiNet	65.90	**66.04**	**1.200**	**0.406**	**73.19**	**72.49**	1.099	**0.623**	**46.35**	**28.69**	**39.86**	28.57	**31.87**	**24.12**	**26.75**	**23.92**
Random Missing Rate $\rho = 0.8$																
MISA	**63.56**	63.16	1.311	**0.321**	63.43	49.95	0.927	0.267	35.33	21.55	22.55	22.60	20.18	16.56	19.17	16.72
Self-MM	59.55	49.98	1.313	0.282	65.07	54.44	0.805	0.259	34.57	23.13	26.25	23.26	21.23	16.53	18.74	16.97
MMIM	58.33	52.46	1.312	0.287	64.97	54.76	0.858	0.269	36.52	23.18	26.95	23.21	20.87	15.92	18.51	16.27
CENET	60.93	54.68	1.314	0.274	65.88	56.80	0.798	0.316	33.95	21.86	24.86	22.56	21.09	17.17	19.94	17.17
TFR-Net	58.49	44.70	1.497	0.155	65.05	54.91	0.807	0.241	**43.04**	23.08	24.68	23.16	23.57	17.73	20.87	18.44
ALMT	60.37	**65.45**	1.310	0.273	68.15	**69.12**	**0.774**	0.231	35.74	22.56	24.02	23.11	20.88	17.26	19.84	17.45
LNLN	62.75	62.56	**1.283**	0.314	68.62	67.81	0.791	0.325	38.85	20.52	27.60	21.70	24.99	17.68	**21.52**	18.60
CaDiNet	61.48	61.21	1.339	0.288	**70.26**	69.07	1.167	**0.623**	42.74	**23.48**	**40.56**	**23.56**	**26.51**	**18.03**	21.22	**18.74**
Random Missing Rate $\rho = 0.9$																
MISA	58.64	56.84	1.369	**0.226**	62.95	48.80	0.941	0.180	31.57	14.92	20.24	17.26	15.54	11.16	14.59	11.71
Self-MM	58.59	46.16	1.353	0.197	63.85	51.32	0.821	0.188	30.63	14.87	20.19	17.21	16.42	12.22	15.39	12.66
MMIM	55.29	47.33	1.357	0.186	63.69	51.26	0.877	0.197	33.26	15.07	20.34	17.46	16.06	11.92	15.73	12.43
CENET	**58.99**	50.01	1.357	0.181	64.14	54.27	0.814	0.254	32.74	15.02	20.29	17.31	15.62	11.35	14.91	11.88
TFR-Net	57.93	43.01	1.534	0.155	63.64	52.02	0.820	0.175	36.28	15.12	20.39	17.41	18.02	12.91	17.01	13.02
ALMT	57.32	**64.92**	1.349	0.205	61.41	63.32	0.810	0.138	30.82	14.82	20.14	17.16	16.86	12.21	16.51	12.60
LNLN	56.50	56.32	1.349	0.202	64.83	61.31	0.820	0.221	34.75	14.97	20.44	17.36	15.94	11.06	16.48	11.50
CaDiNet	58.78	59.33	**1.325**	0.204	**66.19**	**67.44**	**0.726**	**0.484**	**37.59**	**15.42**	**20.74**	**17.76**	**20.91**	**13.13**	**20.22**	**13.32**

Table 11: Results with complete multimodal observations, where CaDiNet remains competitive while preserving its advantage over MISA.

Method	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
MISA	82.78	82.83	0.771	0.777	85.28	85.10	0.552	0.759	57.47	45.04	43.95	46.31	58.13	52.18	48.69	49.98
Self-MM	85.22	85.19	**0.720**	0.790	85.34	85.11	**0.531**	0.764	60.62	**50.33**	50.94	48.00	55.32	52.20	46.70	51.71
MMIM	83.43	83.43	0.744	0.778	83.53	83.39	0.576	0.724	59.24	44.51	44.54	44.55	49.08	49.44	45.94	47.47
CENET	83.08	83.06	0.748	0.785	85.49	85.41	**0.531**	**0.770**	56.98	44.47	43.39	44.40	57.25	46.67	**55.20**	45.75
TFR-Net	83.64	83.57	0.805	0.760	84.96	84.71	0.550	0.745	60.03	46.06	45.04	46.76	48.64	49.52	44.94	49.79
ALMT	84.91	85.01	0.752	0.768	85.62	85.69	0.542	0.752	51.46	40.94	51.86	40.16	52.07	48.54	46.21	46.56
UniMF	82.91	83.12	0.785	0.779	83.47	83.58	0.581	0.728	57.92	44.78	43.60	45.53	56.86	51.00	47.20	47.95
LNLN	84.25	84.61	0.751	0.778	84.14	84.53	0.572	0.735	61.50	47.82	46.31	47.54	54.66	43.68	45.99	43.27
CaDiNet	**85.67**	**85.59**	0.731	**0.796**	**86.32**	**86.25**	0.533	0.769	**62.76**	49.93	**52.92**	**48.98**	**59.32**	**53.27**	54.80	**52.77**

G Detailed Results under Modality-level Missingness

This appendix provides detailed results under different modality-availability patterns. For clearer analysis, we group the results into three settings: complete multimodal observations, one missing modality, and two missing modalities. The averaged results over all input-availability patterns are reported in Table 2 of the main paper.

Complete multimodal observations. As shown in Table 11, the advantage of CaDiNet over strong complete-input baselines is relatively modest, which is expected because the proposed framework is mainly designed for incomplete observations. Nevertheless, CaDiNet still achieves strong overall performance and consistently outperforms MISA (Hazarika et al., 2020), which uses a similar shared–private decomposition. This suggests that further refining the disentangled spaces with complementary evidence can improve multimodal representations even when all modalities are available.

One missing modality. Table 12 shows that CaDiNet achieves strong overall performance under all three single-modality missing settings. The effect of missingness, however, differs substantially across modalities. When language is unavailable (VA), performance drops sharply on all four datasets, whereas removing either the visual (LA) or acoustic (LV) modality causes only limited degradation and occasionally yields small metric fluctuations. This confirms the dominant role of language in these tasks (Zhang et al., 2023; 2024b), with visual and acoustic signals mainly providing complementary information. Moreover, combining visual and acoustic inputs without language brings only limited gains over using either non-linguistic modality alone, particularly on MELD-DA and MIntRec2.0. This suggests that the two auxiliary modalities contain insufficient high-level semantic information to compensate for the absence of language, even when used jointly.

Table 12: Results with one missing modality show consistent advantages of CaDiNet across different missing patterns.

Method	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
VA																
MISA	52.18	48.09	1.422	0.171	62.85	48.51	0.946	0.128	31.60	11.30	13.28	13.05	21.70	14.10	14.69	15.21
Self-MM	57.77	42.31	1.381	0.129	63.03	49.22	0.832	0.152	30.29	10.90	12.90	12.63	22.58	14.36	15.15	15.54
MMIM	54.67	46.95	1.363	0.153	62.71	54.70	0.886	0.195	33.11	12.49	13.56	14.39	22.76	14.47	15.06	15.67
CENET	57.67	42.26	1.387	0.118	63.23	52.41	0.821	0.238	31.56	12.41	13.65	14.06	24.69	16.53	17.27	17.93
TFR-Net	57.82	42.43	1.563	0.167	62.85	48.88	0.829	0.163	36.52	12.82	15.31	15.06	26.08	16.87	17.55	18.17
ALMT	56.40	**66.99**	1.394	0.139	54.28	69.52	0.867	0.071	33.48	12.82	15.29	15.09	23.07	14.54	15.56	15.75
UniMF	55.79	55.79	1.373	0.156	64.28	61.79	0.813	0.200	35.99	11.53	13.32	13.95	25.69	16.85	17.55	18.18
LNLN	52.18	58.89	1.427	0.075	62.85	**77.19**	0.847	0.156	34.69	12.16	13.54	14.69	25.40	16.61	17.30	18.06
CaDiNet	**60.67**	59.85	**1.351**	**0.194**	**66.99**	63.54	**0.796**	**0.276**	**36.94**	**13.08**	**15.60**	**15.40**	**26.46**	**17.19**	**17.91**	**18.55**
LA																
MISA	82.37	82.43	0.777	0.777	85.25	85.11	0.550	0.757	53.15	47.57	53.84	41.77	51.84	42.27	46.05	42.05
Self-MM	85.11	85.09	**0.722**	**0.789**	85.20	84.95	0.533	**0.761**	55.94	46.75	54.30	41.56	49.31	46.80	50.31	47.02
MMIM	83.54	83.52	0.740	0.778	83.71	83.53	0.579	0.722	55.06	44.03	50.95	41.72	54.39	50.50	51.63	50.25
CENET	83.08	83.06	0.748	0.785	84.99	85.03	0.544	**0.761**	54.27	47.14	**60.33**	45.49	48.45	47.57	47.52	48.52
TFR-Net	83.54	83.41	0.799	0.760	83.97	83.93	0.577	0.726	59.89	45.23	52.27	43.43	53.88	44.86	49.62	44.42
ALMT	84.81	84.81	0.863	0.768	66.81	71.72	0.559	0.747	57.40	43.94	52.00	38.29	52.84	49.40	50.35	49.05
UniMF	82.77	82.76	0.849	0.730	81.10	80.51	0.607	0.675	51.22	39.80	49.14	38.06	49.16	44.34	46.86	44.73
LNLN	84.91	85.17	0.759	0.772	84.09	84.43	0.577	0.729	62.08	48.20	59.45	46.05	58.72	51.18	55.74	50.91
CaDiNet	**85.52**	**85.48**	0.723	**0.789**	**86.12**	**86.09**	**0.531**	0.753	**62.46**	**48.54**	59.93	**46.42**	**59.08**	**51.55**	**56.16**	**51.28**
LV																
MISA	83.74	83.72	0.762	0.776	84.94	84.62	0.561	0.759	51.79	**52.27**	53.31	42.32	49.17	44.17	46.16	44.68
Self-MM	85.22	85.19	**0.720**	0.790	85.31	85.06	0.532	0.763	54.79	46.55	57.28	42.22	51.75	48.25	**56.42**	49.01
MMIM	83.08	83.00	0.744	0.777	83.72	83.60	0.572	0.726	58.55	47.04	54.10	43.39	55.13	50.71	51.53	51.87
CENET	83.08	83.06	0.750	0.785	85.37	85.35	0.531	0.770	53.45	43.91	57.64	42.30	58.76	52.59	55.16	52.12
TFR-Net	82.72	82.71	0.821	0.759	84.95	84.69	0.550	0.745	61.90	45.15	58.81	42.82	57.41	50.64	52.26	50.58
ALMT	84.60	84.81	0.865	0.767	67.13	71.76	0.535	0.556	62.72	51.43	**65.23**	49.42	59.50	53.20	55.84	52.74
UniMF	81.10	81.08	0.918	0.697	82.10	81.93	0.605	0.682	54.85	50.83	59.98	48.87	49.44	44.00	47.09	43.61
LNLN	84.86	85.12	0.760	0.772	84.39	84.71	0.610	0.736	59.62	44.44	57.56	40.89	57.96	51.37	54.17	51.98
CaDiNet	**86.59**	**86.41**	0.721	**0.792**	**86.30**	**86.18**	**0.529**	**0.772**	**63.16**	51.87	65.01	**49.87**	**59.96**	**53.66**	56.29	**53.18**

Two missing modalities. The results in Table 13 further reveal a strong modality asymmetry. When only language is retained, the performance of most methods decreases only slightly compared with complete multimodal inputs. In contrast, retaining only visual or acoustic observations causes a substantial performance drop, particularly on the fine-grained MELD-DA and MIntRec2.0 datasets. With visual input alone, the best accuracies are only around 35% on MELD-DA and 17% on MIntRec2.0, showing that facial and visual cues alone provide limited information for distinguishing semantically similar dialogue acts and intentions.

Acoustic input alone performs slightly better than visual input on these two datasets, but remains far below the text-only setting. One possible reason is that the Librosa-based acoustic features mainly encode low-level properties such as spectrum, pitch, energy, and rhythm (McFee et al., 2015; Zadeh et al., 2016). These cues characterize how an utterance is spoken, but contain limited information about its lexical content and high-level contextual semantics. Consequently, they are useful for capturing prosody and affective states but are insufficient for recovering complex semantic distinctions when language is absent.

Overall, these results confirm that language carries most of the task-relevant semantics in the considered benchmarks, while visual and acoustic modalities mainly provide complementary cues. CaDiNet nevertheless achieves strong overall performance under severe modality loss, indicating that its calibration and evidence-guided refinement can better exploit the limited information remaining in the available modalities.

H Mechanistic Analysis of Missingness-Induced Representation Shifts

In the main paper, we argue that incomplete observations induce representation shifts from their complete counterparts, and that such shifts may accumulate during subsequent cross-modal interaction. To examine this phenomenon more directly, we conduct a layer-wise analysis using MAG-BERT (Rahman et al., 2020) as the backbone. MAG-BERT contains a standard BERT stack with

Table 13: Results with two missing modalities show the strong dependence of multimodal language understanding on textual information.

Method	CMU-MOSI				CMU-MOSEI				MELD-DA				MIntRec2.0			
	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	MAE(↓)	Corr(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)	Acc(↑)	m-F1(↑)	m-P(↑)	m-R(↑)
V																
MISA	55.29	49.53	1.419	0.098	62.85	48.51	0.961	0.121	30.62	8.94	6.80	12.17	13.71	3.84	3.41	6.22
Self-MM	57.77	42.31	**1.381**	0.130	63.01	49.18	0.831	0.144	34.62	9.48	7.68	13.62	16.31	4.41	3.90	7.26
MMIM	52.24	39.57	1.389	0.024	62.31	54.32	0.886	0.182	32.39	9.46	7.80	13.17	14.73	3.90	3.45	6.42
CENET	57.67	42.26	1.387	0.110	60.69	51.22	0.823	0.236	30.75	9.61	7.40	13.76	16.26	4.31	3.92	7.07
TFR-Net	58.03	43.37	1.554	0.107	62.81	48.75	0.829	0.161	**35.32**	9.55	7.72	13.70	16.38	4.45	3.94	7.30
ALMT	56.35	**66.89**	1.397	0.104	54.28	69.52	0.870	0.091	30.93	9.36	7.01	13.65	15.06	3.99	3.68	6.56
UniMF	54.12	53.05	1.393	0.137	64.61	62.40	**0.815**	0.195	30.50	9.46	7.22	13.32	15.89	4.23	3.75	6.78
LNLN	52.18	58.89	1.427	0.072	62.85	**77.19**	0.900	0.145	32.75	9.10	6.75	13.14	16.39	4.59	4.07	**7.99**
CaDiNet	**61.28**	60.21	1.396	**0.163**	**65.66**	61.37	0.821	**0.246**	35.14	**9.81**	**7.96**	**14.04**	**16.72**	**4.68**	**4.15**	7.59
A																
MISA	48.98	41.79	1.427	0.169	62.85	48.51	0.936	0.105	31.44	13.91	19.72	14.25	20.83	12.31	12.23	13.95
Self-MM	**57.77**	42.31	1.386	0.072	62.85	48.51	0.838	0.101	32.87	13.69	18.39	14.48	22.61	12.82	12.78	14.74
MMIM	53.05	43.80	**1.366**	0.155	62.88	48.91	0.891	0.143	33.16	14.64	**21.13**	15.34	21.29	12.47	12.29	14.20
CENET	57.67	42.26	1.387	0.118	61.78	51.87	0.838	0.112	34.45	13.86	20.03	17.03	23.39	13.36	13.26	15.33
TFR-Net	**57.77**	42.31	1.486	0.154	62.47	51.16	0.839	0.039	37.49	14.35	19.51	15.69	23.74	13.56	13.79	15.49
ALMT	56.45	**67.09**	1.394	0.136	54.28	69.52	0.874	-0.086	37.86	14.54	20.25	16.92	24.70	**15.12**	14.50	16.45
UniMF	54.88	51.67	1.473	0.230	62.60	51.15	0.836	0.036	37.57	12.25	17.00	14.40	23.61	14.70	14.61	16.57
LNLN	52.18	58.89	1.427	0.075	62.85	**77.19**	0.853	0.052	37.95	14.61	20.34	**17.52**	24.79	14.66	14.57	16.54
CaDiNet	57.47	56.81	1.418	**0.256**	**64.78**	60.64	**0.814**	**0.219**	**38.34**	**14.94**	20.73	17.38	**25.14**	15.00	**14.91**	**16.91**
L																
MISA	83.28	83.29	0.768	0.776	84.88	84.64	0.558	0.756	51.75	41.72	44.50	41.70	48.48	42.12	43.66	42.51
Self-MM	85.06	85.04	0.722	0.789	85.32	85.06	0.535	0.761	62.65	50.42	53.81	49.55	58.66	50.92	52.80	51.40
MMIM	83.48	83.42	0.748	0.777	83.79	83.61	0.579	0.722	60.29	42.31	47.90	42.09	55.81	50.25	50.52	50.04
CENET	83.08	83.06	0.750	0.785	84.64	84.71	0.545	0.760	55.00	42.64	47.36	40.98	56.92	50.33	50.84	50.80
TFR-Net	83.08	83.01	0.838	0.758	83.97	83.91	0.578	0.726	60.65	44.72	48.25	43.63	57.94	44.73	43.73	45.52
ALMT	84.55	84.76	0.864	0.767	67.14	71.77	0.560	0.748	57.28	46.56	53.18	45.31	48.87	48.37	49.16	48.57
UniMF	82.32	82.28	0.908	0.699	81.16	81.05	0.609	0.669	54.78	49.86	51.68	48.97	48.69	44.52	47.18	44.26
LNLN	84.86	85.12	0.760	0.772	84.10	84.39	0.609	0.730	61.85	44.91	48.59	43.46	57.34	**51.76**	44.37	43.40
CaDiNet	**85.82**	**85.78**	**0.721**	**0.798**	**85.57**	**85.77**	**0.531**	**0.762**	**63.11**	**50.88**	**54.27**	**49.97**	**59.12**	51.36	**53.24**	**51.84**

a clear layer-wise structure, which allows us to trace how the discrepancy between complete and incomplete inputs evolves from the embedding layer to the final semantic representation.

H.1 Experimental Setup

We conduct the analysis on CMU-MOSI and construct incomplete observations with a token-level missing ratio of $\rho = 0.5$. For the same sample, we separately feed its complete and incomplete versions into the model and measure the Euclidean distance between their corresponding representations at each BERT layer.

We focus on the representation of the `[CLS]` token, since it gradually aggregates contextual semantics through the BERT layers and serves as the semantic representation used for downstream prediction. Specifically, for layer l, we compute

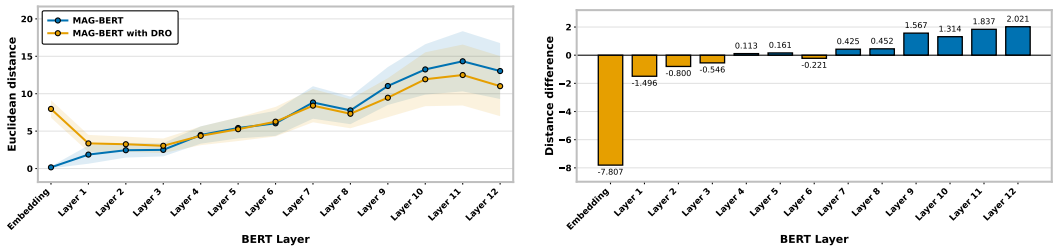
$$d^{(l)} = \left\| \mathbf{z}_{\text{cls}}^{(l),\text{comp}} - \mathbf{z}_{\text{cls}}^{(l),\text{miss}} \right\|_2, \tag{17}$$

where $\mathbf{z}_{\text{cls}}^{(l),\text{comp}}$ and $\mathbf{z}_{\text{cls}}^{(l),\text{miss}}$ denote the `[CLS]` representations obtained from the complete and incomplete observations, respectively. The reported values are averaged over the test samples.

Since the original MAG-BERT fusion module operates on token sequences, whereas DRO is designed to calibrate modality-level semantic representations, we make a minimal adaptation when introducing DRO into MAG-BERT. We first average the token representations of each modality to obtain its modality-level representation and apply DRO to it. The resulting residual correction

$$\Delta \mathbf{h}_m = \rho_m \left(\boldsymbol{\pi}_m + \mathcal{A}_m(\mathbf{c}_m) \right) \tag{18}$$

is then injected only into the `[CLS]` token. In this way, the original token-level interaction structure of MAG-BERT remains unchanged, while the effect of representation calibration can be observed directly from the semantic token.



(a) (b)

Figure 9: Layer-wise analysis of missingness-induced representation shifts on CMU-MOSI ($\rho = 0.5$). (a) Euclidean distance between complete and incomplete [CLS] representations. (b) Distance reduction brought by DRO, showing its increasing calibration effect in deeper layers.

Table 14: Performance of MAG-BERT on CMU-MOSI under complete and incomplete observations. For the incomplete setting, the token-level missing ratio is $\rho = 0.5$. Δ denotes the improvement obtained by introducing DRO.

Version	Has0 Acc-2	Has0 F1	Non0 Acc-2	Non0 F1	Acc-5	Acc-7	MAE	Corr
Complete observation								
MAG-BERT	82.22 ± 0.63	82.12 ± 0.80	84.35 ± 0.39	84.33 ± 0.39	40.81 ± 6.09	37.22 ± 4.30	0.840 ± 0.045	0.767 ± 0.014
Incomplete observation with missing ratio $\rho = 0.5$								
MAG-BERT	66.91 ± 2.41	66.59 ± 3.08	67.23 ± 2.92	67.02 ± 3.57	29.35 ± 3.45	27.16 ± 2.58	1.179 ± 0.048	0.490 ± 0.042
MAG-BERT with DRO	69.05 ± 1.50	69.11 ± 1.50	69.72 ± 1.88	69.88 ± 1.88	32.95 ± 1.39	29.84 ± 1.35	1.152 ± 0.040	0.491 ± 0.029
Δ	$\mathbf{+2.14}$	$\mathbf{+2.52}$	$\mathbf{+2.49}$	$\mathbf{+2.86}$	$\mathbf{+3.60}$	$\mathbf{+2.68}$	$\mathbf{-0.027}$	$\mathbf{+0.001}$

H.2 Layer-wise Analysis of Representation Shifts

Figure 9(a) shows the layer-wise Euclidean distance between the complete and incomplete [CLS] representations. For the original MAG-BERT, the distance is close to zero at the embedding layer because the initial [CLS] token is identical for complete and incomplete inputs and has not yet interacted with the corrupted tokens. As the representation propagates through the BERT layers, the distance generally increases and becomes substantially larger in deeper layers. This observation indicates that the effect of missing inputs is progressively propagated into the high-level semantic representation, providing direct evidence for the missingness-induced representation shifts discussed in the main paper.

For MAG-BERT with DRO, a relatively large distance appears at the embedding layer. This is expected because DRO explicitly injects the residual correction $\rho_m(\boldsymbol{\pi}_m + \mathcal{A}_m(\mathbf{c}_m))$ into the [CLS] token of the incomplete input, whereas the original [CLS] token is initially unaffected by missingness. More importantly, this initial discrepancy is gradually reduced through subsequent contextual modeling.

To show this effect more clearly, Fig. 9(b) reports $d_{\text{MAG-BERT}} - d_{\text{MAG-BERT+DRO}}$, where a positive value indicates a smaller complete–incomplete distance after calibration. The difference changes from negative in the early layers to positive in the deeper layers and reaches 2.021 at Layer 12. This trend suggests that DRO does not simply minimize the initial representation distance; instead, it provides a corrective offset before semantic interaction, which progressively mitigates the propagation of missingness-induced shifts.

H.3 Effect on Downstream Prediction

We further examine whether the representation-level improvement translates into better downstream predictions. Table 14 reports the results of MAG-BERT on CMU-MOSI under complete observations and under incomplete observations with $\rho = 0.5$. Has0 evaluates negative versus non-negative sentiment, whereas Non0 excludes neutral samples and evaluates negative versus positive sentiment. Acc-5 and Acc-7 denote five- and seven-class accuracy, respectively.

Missing observations cause a clear performance degradation in the original MAG-BERT. After introducing DRO, all six classification metrics improve, with an average absolute gain of approximately

2.72 percentage points. MAE is reduced from 1.179 to 1.152, while Corr increases slightly from 0.490 to 0.491. The improvement is particularly clear on Acc-5 and Acc-7, which increase by 3.60 and 2.68 percentage points, respectively.

Together with the layer-wise analysis, these results provide two complementary observations. First, incomplete observations introduce discrepancies that are progressively propagated into high-level semantic representations during cross-modal interaction. Second, calibrating the representation before such interaction changes its subsequent propagation trajectory and reduces the final complete–incomplete representation gap. The corresponding improvement in downstream performance further supports the motivation of the proposed "calibrate-then-disentangle" paradigm: mitigating the missingness-induced shift before extensive interaction provides a cleaner representation for subsequent multimodal modeling.