T^{2}DR: A Two-Tier Deficiency-Resistant Framework for Incomplete Multimodal Learning

Han Lin, Xiu Tang[*], Huan Li, Wenxue Cao, Sai Wu, Chang Yao, Lidan Shou, Gang Chen
Zhejiang University,
Hangzhou High-Tech Zone (Binjiang) Blockchain and Data Security Research Institute,
Zhejiang Key Laboratory of Big Data Intelligent Computing.
{22351123, tangxiu, lihuan.cs, 22451028, wusai, changy, should, cg}@zju.edu.cn

Abstract

Multimodal learning is garnering significant attention for its capacity to represent diverse human perceptions (e.g., linguistic, acoustic, and visual signals), achieving more natural and intuitive interactions with technology. However, the frequent occurrence of incomplete data, either within a single modality (intra-modality) or across different modalities (inter-modality), presents substantial challenges in reliable semantic interpretation and model reasoning. Furthermore, there is currently no robust representation learning mechanism capable of managing both intra-modality and inter-modality real-data deficiencies. To address this challenge, we present T^{2}DR, a two-tier deficiency-resistant framework for incomplete multimodal learning, which comprises two main modules: (1) Intra-Modal Deficiency-Resistant module (IADR): To address fine-grained deficiencies, we introduce Intra-Attn to focus on the available data while avoiding excessive suppression of the missing regions. (2) Inter-Modal Deficiency-Resistant module (IEDR): To handle coarse-grained deficiencies, we propose the shared feature prediction (SFP) to leverage cross-modal shared features for preliminary data imputation. Subsequently, we apply Inter-Attn to allocate appropriate attention to each modality based on the results from the capability-aware scorer (CAS). Extensive experiments are performed on two well-known multimodal benchmarks, CMU-MOSI and CMU-MOSEI, across various missing scenarios for sentiment analysis. Experimental results show that T^{2}DR significantly outperforms the SOTA models. Code is available at `https://github.com/LH019/T2DR`.

1 Introduction

In modern data analysis, the significance of multimodal data is increasingly prominent. Unlike unimodal data, multimodal data can capture diverse information from heterogeneous sources and leverage the complementary of different modalities for more sufficient information (Wei et al., 2023b). Nevertheless, the integration process is often compromised by the presence of incomplete data (Guo et al., 2024). This incomplete data may present as fine-grained (intra-modality), shown by slight variations or gaps within a single modality (Zhang et al., 2019), typically arising from interruptions in data collection or improper data management. It can also appear as coarse-grained (inter-modality) (Zhao et al., 2021a), manifested as significant omissions across different modalities, commonly due to unavailable data sources or failures to acquire multimodal data synchronously. Existing methods for handling both intra-modal and inter-modal incomplete data (Yuan et al., 2023, 2024) typically rely on randomly masking features derived from complete data to simulate intra-modal incompleteness, which fails to faithfully represent the real-world scenarios. Therefore, addressing the issue of fine-grained and coarse-grained deficiencies is crucial for improving the precision and reliability of multimodal learning.

Existing methods for addressing incomplete multimodal learning are divided into three main categories: generative methods (Liu et al., 2023; Xu et al., 2019; Tang and Liu, 2022), multimodal joint learning (Qu et al., 2024; Liu et al., 2024b; Zhao et al., 2021b), and knowledge distillation (Xing et al., 2022; Poklukar et al., 2022a). Generative methods such as VIGAN (Shang et al., 2017) use generative adversarial networks alongside denoising autoencoders to recover and refine incomplete modalities. Multimodal joint learning methods, exemplified by DrFuse (Yao et al., 2024), extract shared information from available modalities to compensate for missing modalities. Knowledge distillation approach MMANet (Wei et al., 2023a) leverages a margin-aware distillation strategy to help the student network improve its comprehension of inter-class relationships, enhancing the final

[*] Corresponding author

Findings of the Association for Computational Linguistics: ACL 2025, pages 8602–8616
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

shared feature representation.

While existing methods have shown promising results in handling missing modalities, they often focus on the quality of the imputed features (Zhao et al., 2021a), with substantial challenges remaining. These include: (1) a lack of dynamic adjustment mechanisms during model training to resist interference from incomplete data, especially for imputed features with lower quality, and (2) a predominant focus on inter-modality incomplete data scenarios, often overlooking the more common intra-modality incomplete data, where the missingness occurs at the raw data level rather than the processed feature level.

To address the above two challenges, we present a novel approach — a **T**wo-**T**ier **D**eficiency-**R**esistant Framework for Incomplete Multimodal Learning, referred to T^2DR, which focuses on deeper learning from valid data and enhancing its resistance to various deficiencies. Our main contributions can be summarized as follows:

- We propose T^2DR to resist deficiency for incomplete multimodal learning. For fine-grained deficiencies, we introduce Intra-Attn to resist noise disruption by dynimically balancing the missing parts' effect with the model's global capacity; for coarse-grained deficiencies, we introduce Inter-Attn to resist low-quality feature damage by dynamically distributing attention to each modality.

- Facing coarse-grained deficiencies, to resist zero-information in missing modalities, we introduce the shared feature prediction (SFP) method, utilizing existing modalities to predict shared information across modalities to fill in the gaps. To resist the negative impact of low-quality predicted features, we employ the capability-aware scorer (CAS), a task-driven modality capability-aware technique, dynamically computing weights for effective supervision in Inter-Attn.

- We conduct comprehensive evaluations of our method across multiple datasets. To investigate fine and coarse-grained incomplete data in multimodal learning, we conduct multimodal sentiment analysis on CMU-MOSI and CMU-MOSEI. To verify its effectiveness in unimodal learning, we also conduct classification tasks using CARER for text, ESC-50 for acoustic, and ImageNet for visual. Our results demonstrate that T^2DR achieves state-of-the-art performance, consistently outperforming existing models and confirming its robustness in handling incomplete multimodal data.

2 Related Work

To tackle deficiencies in multimodal learning, we enhance the model's adaptability by using *Missing Modality Prediction* for a robust end-to-end system, *Weight Allocation* to perform dynamic supervision within each data entry, and *Modality Fusion* to leverage modality complementarity for richer semantic features.

Missing Modality Prediction. Missing modality prediction leverages the intrinsic dependencies across modalities to fill in missing entries. Current dominant methods typically design a unified representation to preserve the expression of shared features, which generally follow three research lines: (1) explicit disentanglement of shared and modality-specific information, utilizing shared information from available modalities for completion (Yao et al., 2024; Wang et al., 2023); (2) direct construction of a unified shared representation, minimizing reliance on specific modalities (Lau et al., 2019; Nawaz et al., 2024); (3) knowledge distillation via teacher networks, learning modality-specific and shared features (Wei et al., 2023a). However, these methods often rely on complex networks and learning objectives to ensure the quality of the unified representation. They lack post-processing steps to alleviate the inevitable impact of missing modalities on feature quality.

Weight Allocation. Weight allocation assigns varying importance to samples or features based on expectations. In sample-based allocation methods, weighted summation (Wan et al., 2023; Ahmadianfar et al., 2022) and weighted cross-entropy loss (Legate et al., 2023; Rezaei-Dastjerdehei et al., 2020) assign proper weight to each sample. In feature-based allocation methods, masked multi-head attention in Transformer (Vaswani et al., 2017) allocates the entire weight on the token before the current position to avoid accessing future information. Furthermore, some methods (Li et al., 2022; Sergio and Lee, 2021) addressing incomplete data allocate zero weight on the missing parts to mitigate the adverse effects of noisy data. While these methods have performed weight adjustment within features, they still remain limited to whether or not the part is allocated weight.

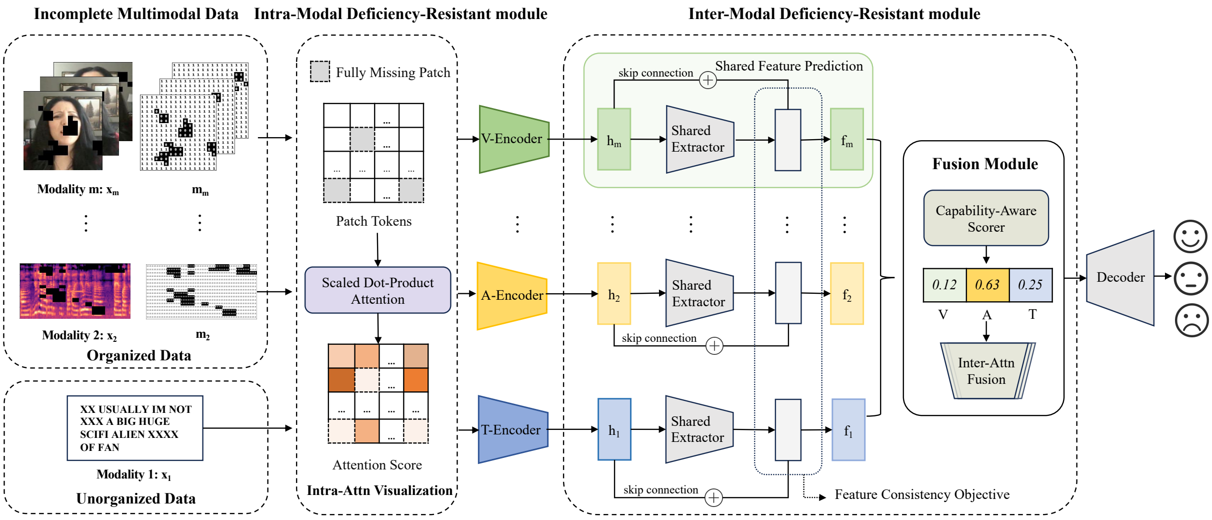


Figure 1: **Processing pipeline of T^{2}DR for multimodal sentiment analysis.** T^{2}DR performs two-stage resistance for incomplete multimodal data: (1) in the fine-grained stage, multimodal data $\{x_1, x_2, \ldots, x_m\}$ are filtered by *Intra-Attn* according to $\{m_1, m_2, \ldots, m_m\}$ and then processed to produce modality-specific features $\{h_1, h_2, \ldots, h_m\}$ in a unified latent space; (2) in the coarse-grained stage, *shared feature prediction* (SFP) enhances these features with richer semantics $\{f_1, f_2, \ldots, f_m\}$. An advanced *fusion module* combines these features through *capability-aware scorer* (CAS) and *Inter-Attn* for reliable prediction.

Modality Fusion. Modality fusion combines heterogeneous data to achieve a comprehensive understanding of the sample. Transformer-based fusion methods often fall into three categories: (1) Fusion Token approaches (Tan and Bansal, 2019; Nagrani et al., 2021) employ bidirectional cross-attention to align and exchange information across modalities, ultimately extracting fusion token embeddings to serve as the fusion representation. However, considering the limited operable steps, this strategy has not been widely applied to multimodal fusion. (2) Co-attention strategies (Feng et al., 2021; Zou et al., 2022; Chen et al., 2021), like FTransUNet (Ma et al., 2024) and MCSAN (Sun et al., 2021), combine self-attention and cross-attention to capture intra-modal and cross-modal information. However, these methods struggle to scale with more than two modalities, as the exponential growth in required computational modules leads to a significant surge in in complexity and resource demands. (3) Concatenation-based Attention (Liu et al., 2024a; Yao et al., 2024), commonly applied in scenarios involving more than two modalities, such as MFM-Att (Fang et al., 2023) and UNITER (Chen et al., 2020), reorganizes tokens or concatenates multimodal features vertically and horizontally to generate fused features. However, these methods still lack targeted solutions for incomplete multimodal data.

3 Method

3.1 Model Overview

In this section, we formulate the problem definition and present an overview of our proposed solution.

Definition. Given m modalities of a data instance $X = \{x_1, x_2, \ldots, x_m\}$ and a set of mask vectors $M = \{m_1, m_2, \ldots, m_m\}$ indicating the data completeness, we train a model T^{2}DR$(X, M) \to Y$. The label Y signifies the corresponding label for the downstream task.

Scheme Overview. Figure 1 illustrates an overview of T^{2}DR architecture, which consists of two modules: Intra-Modal Deficiency-Resistant module (IADR) and Inter-Modal Deficiency-Resistant module (IEDR), achieving robustness and adaptability in incomplete multimodal learning. We focus our research on three modalities: linguistic, visual, and acoustic. The general processing pipeline is outlined as follows.

For fine-grained deficiencies, T^{2}DR adopts *Intra-Attn$_i$* to filter x_i, allowing it to mainly focus on the non-missing data segments corresponding to where m_i equals 1. Then, specific pre-trained encoder $E_i(\cdot)$ and unified mappings $U_i(\cdot)$ map the encoded features h_i into a unified feature space, which can be expressed as follows:

$$h_i = U_i(E_i(\text{Intra-Attn}_i(x_i, m_i))). \qquad (1)$$

For coarse-grained deficiencies, the *Shared Feature Prediction* (SFP) utilizes the shared extractor $S_i(\cdot)$ to capture cross-modal shared information for imputing the missing modality and builds enhanced semantic features f_i via a residual connection:

$$f_i = S_i(h_i) + h_i. \tag{2}$$

Then, *Inter-Attn* integrates the concatenated features $(f_1, f_2, \ldots, f_m)$ with the weights from *Capability-Aware Scorer* (CAS). The fused feature f is acquired as follows:

$$\begin{aligned} f = \text{Inter-Attn}([f_1, f_2, \ldots, f_m], \\ \text{CAS}([f_1, f_2, \ldots, f_m])). \end{aligned} \tag{3}$$

The fused feature f is then passed to the decoder to obtain the final predicted label $\hat{y}$:

$$\hat{y} = dec(f). \tag{4}$$

3.2 Intra-Modal Deficiency-Resistant Module

Motivation. In the context of incomplete images, traditional visual methods (Li et al., 2022; Sergio and Lee, 2021) often employ masked multi-head attention to ignore missing parts, thereby mitigating the impact of noise. However, this approach tends to excessively depend on the available information while completely disregarding the potential value of the missing parts. This local bias may limit the model's ability to capture global context. To validate this hypothesis, we conducted comprehensive experiments on ViT (Dosovitskiy et al., 2020), analyzing how the index "attention distance," which quantifies how far each token can effectively attend to others, varies with different degrees of data incompleteness. Our results show that as the missing rate increases, the attention distance significantly decreases, indicating a degradation in the model's global capability. Detailed experimental results and theoretical analysis can be found in Appendix A.1. To address this issue, we propose *Intra-Attn* to smooth the attention of missing parts and enhance the model's generalization ability.

Data Preprocessing. The digitized form of a modality x_i is typically represented as either a one-dimensional vector, such as a token embedding for text, or a two-dimensional matrix, like a pixelated image for visual or a pixelated spectrogram for acoustic. However, if matrices are directly input into Transformer-based models, each pixel must compute self-attention with all other pixels, resulting in quadratic growth in computational complexity (Dosovitskiy et al., 2020). Hence, preprocessing is essential for matrices and simpler vectors can skip this step.

In detail, matrices need to be divided into patches of size *window_size* × *window_size*, and then flattened into embeddings. To enhance global feature capture across patches and maintain data coherence, a window-shifting operation (Liu et al., 2021) is performed before the division:

$$\begin{aligned} x' &= \text{roll}(x, -s, \text{dims} = (1, 2)), \\ m' &= \text{roll}(m, -s, \text{dims} = (1, 2)), \end{aligned} \tag{5}$$

where roll shifts the elements of the matrix x and m by s steps along the specified dimensions, x represents the complete image, m signifies $x's$ mask information, and s is the *shifted_size*, which is equal to *window_size*/2.

Intra-Attn Processing. Since the encoder plays a crucial role in extracting semantic features, it is essential to minimize the interference of intra-modality incomplete data on its input. To filter the noisy x_i, we focus predominant attention on the valid non-missing data. For the preservation of the model's global capacity, we introduce compensation components for missing parts and apply regulation coefficient ϵ to balance the competitive dynamics between missing and available parts, which is represented as:

$$\text{Intra\_Attn}(x_i) = \sigma\Big((1-\epsilon)\cdot \frac{\mathbf{QK}^\top + \mathbf{M}}{\sqrt{d_k}} + \epsilon \cdot \mathbf{U} \Big) \cdot \mathbf{V}. \tag{6}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are derived from linear transformations of x_i; $\sigma(\cdot)$ represents the Softmax operation; $\frac{1}{\sqrt{d_k}}$ serves as a scaling factor to stabilize the dot product.

The compensation component $\mathbf{U}$ is equivalent to $\frac{\lambda}{\sqrt{d_k}}$, where λ is a trainable parameter that reflects the influence of global information enhancement. The matrix $\mathbf{M}$ is derived from the mask m, where $\mathbf{M}$ is set to $-\infty$ if $m = 0$ (indicating a missing token) and 0 otherwise. It ensures that when a token is missing, the attention value after the compensation components and softmax operation approaches zero. If x_i is a matrix, a missing token indicates a completely missing patch.

3.3 Inter-Modal Deficiency-Resistant Module

Shared Feature Prediction. When encountering inter-modal incomplete data, if T^{2}DR still adopts

Intra-Attn to resist deficiency, it will result in a significant data loss, severely impairing the learning of associated modules. Therefore, a *Shared Feature Prediction* is proposed to extract common features that are relevant to the downstream tasks, while filtering out the redundant information specific to each modality. Specifically, we adopt the Kullback-Leibler (KL) divergence (Van Erven and Harremos, 2014) to ensure Feature Consistency (FC) of available modalities, expressed as follows:

$$\mathcal{L}_{\mathcal{FC}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i,j=1}^{m} \mathrm{KL}\left[\sigma\left(S(h_i^{(b)})\right), \sigma\left(S(h_j^{(b)})\right)\right], \tag{7}$$

where $\sigma(\cdot)$ refers to the softmax operation, b denotes the samples of the b-th batch and i, j represent the corresponding modalities. If the t-th modality is missing, then $i, j \in \{1, \ldots, t-1, t+1, \ldots, m\}$. When the KL divergence is minimized, it indicates that the shared features have been successfully extracted. If the m_i indicates the x_i entirely misses, its features can be replaced by the shared features from available modalities:

$$f_i^{(k)} = \frac{1}{m-1} \cdot \sum_{j=1, j\neq i}^{m} S(h_j^{(k)}). \tag{8}$$

Capability-Aware Scorer. The quality of modality feature f_i is inherently unquantifiable and varies across different downstream tasks. To address this issue, we propose a task-driven quality evaluation mechanism that computes pseudo-ground-truth quality, guiding the *Capability-Aware Scorer* (CAS) to learn feature-specific weight allocation capabilities.

To ensure that the final features retain more task-specific information, we introduce downstream task losses to recognize the capability of f_i. As shown in Figure 2, to achieve precise evaluations of the i-th modality's quality, the input is fed into the decoder by retaining only f_i and setting other modality features to zero during concatenation. The pseudo-quality of f_i is then determined based on the resulting loss:

$$\frac{1}{q_i^{(k)}} = \mathrm{CE}(\mathrm{dec}([0, \ldots, f_i^{(k)}, \ldots, 0]), Y^{(k)}) + \epsilon, \tag{9}$$

where ϵ is a small constant value, typically set to 1e-8, used to avoid division by zero errors, $[0, \ldots, f_i^{(k)}, \ldots, 0]$ represents the concatenation of $f_i^{(k)}$ with $(m-1)$ zeros, and CE denotes the cross-entropy loss. A smaller loss value indicates that i-th modality possesses stronger predictive capability and higher quality of f_i. Then, the pseudo-weight w is computed as follows:

$$w_i^{(k)} = \frac{q_i^{(k)}}{\sum_{j=1}^{m} q_j^{(k)}}, \tag{10}$$

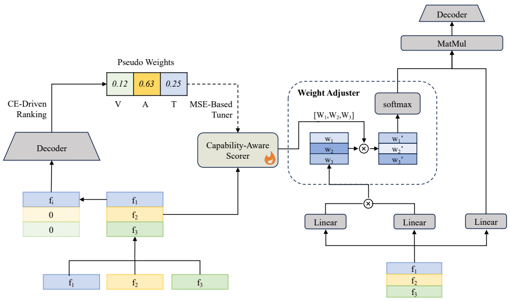


Figure 2: **Fusion module internals.** The feature vector $[0, \ldots, f_i, \ldots, 0]$ is input into the decoder to compute pseudo weights w_i for training the *Capability-Aware Scorer* (CAS). Then W_i is utilized to dynamically adjust the modality fusion.

To ensure precise and task-specific capability perception, an MSE-Based Tuner (MBT) is employed to optimize the CAS module:

$$\begin{aligned}\mathcal{L}_{\mathcal{MBT}} = \sum_{k=1}^{n} \mathrm{MSE}\Big\{&\left[w_{=1}^{(k)}, \ldots, w_m^{(k)}\right], \\ &\mathrm{CAS}\left(\left[f_1^{(k)}, \ldots, f_m^{(k)}\right]\right)\Big\}.\end{aligned} \tag{11}$$

where MSE means Mean Squared Error, which measures the average squared differences between observed and predicted values, ensuring the precision of CAS's predictions.

Inter-Attn Processing. As shown in Figure 2, the Weight Adjuster fuses the features f_i with respective supervision W_i, generated by the CAS. To ensure that the absolute value of the attention scores is a larger value after multiplication with w_i^* and a smaller value after division, the normalized weight W_i is preprocessed as follows:

$$w_i^* = (1 + W_i) \cdot \lambda, \tag{12}$$

where λ is a trainable parameter that reflects the influence of W_i on the final result.

Intra_Attn impairs the final weights assigned to the mask part by using $\mathbf{M}$. However, the addition limits the range of adjustments, making it effective only for setting attention to the minimal near-zero or maximum near-one. Thus, Inter_Attn adopts multiplication to reasonably allocate modality weights:

$$\text{Inter-Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top \cdot \mathbf{M}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{13}$$

$$\mathbf{M} = \begin{cases} w^*, & \text{if } \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \geq 0, \\ \frac{1}{w^*}, & \text{otherwise.} \end{cases} \tag{14}$$

To achieve higher attention with greater weights (given $w^* > 1$), positive values are amplified while negative values are reduced. It ensures that high-quality modality features f_i are assigned higher weights W_i and receive greater attention in the model, maximizing the utilization of the high-quality data.

4 Experimental Setup

Benchmark Datasets. We evaluate T^{2}DR's performance with two multimodal datasets, each containing text, acoustic, and visual modalities, and three unimodal datasets, each focused on one modality: text, acoustic, or visual.

For multimodal datasets, CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018) are sourced from YouTube videos, annotated with sentiment intensity scores from -3 (highly negative) to +3 (highly positive), covering positive, negative, and neutral sentiments. Notably, CMU-MOSEI is nearly ten times larger than CMU-MOSI, offering a more extensive range of speakers and topics.

For unimodal datasets, we use CARER (Saravia et al., 2018) for text, ESC-50 (Srivastava and Sharma, 2024) for acoustic, and ImageNet (Deng et al., 2009) for visual. CARER targets emotion recognition in text, ESC-50 supports environmental sound classification, and ImageNet focuses on image classification across diverse objects and scenes.

Dataset Preprocessing. For unimodal datasets, we simulate scenarios of intra-modality incomplete data through specific preprocessing techniques. For text data, missing data often occurs at the character or word level in real-world scenarios, so we randomly convert original tokens to [UNK] tokens. For visual data, we set all three color channels of a pixel to zero, resulting in a black pixel. For acoustic data, which is typically input as a spectrogram (essentially an image), so we adopt the same simulation technique as for visual data.

For multimodal datasets, existing methods commonly use CMU's preprocessed encoded data to simplify processing (Sun et al., 2022; Zhao et al., 2021c; Poklukar et al., 2022b), which determines the completeness of its intra-modality data. Therefore, we recode the raw data, ensuring that our experiments can cover both intra-modal and inter-modal incomplete data. Specifically, the data after intra-modal incomplete data processing is then input into the corresponding encoder: BERT (Devlin et al., 2018) for text data, OpenFace (Baltrušaitis et al., 2016) for visual data, and OpenSmile (Eyben et al., 2010) for acoustic data. For inter-modal incomplete data, we randomly select a modality from certain samples and set all its encoded features to zero to simulate complete modality loss.

Metrics. We adopt accuracy (Acc) and weighted F1 score as performance metrics, with the latter mitigating the impact of class imbalance in the MOSI and MOSEI. The reported results represent the averages obtained from five repeated experiments.

Baselines. For mixed coarse-grained and fine-grained incomplete data scenarios, we compared T^{2}DR against four reproducible state-of-the-art methods: CubeMLP (Sun et al., 2022) for complete modality scenarios, and RedCore (Sun et al., 2024), MMIN (Zhao et al., 2021c), and GMC (Poklukar et al., 2022b) for incomplete data scenarios.

For fine-grained incomplete data scenarios, we applied two classic methods for each modality to conduct classification tasks: TextCNN (Gong and Ji, 2018) and FastText (Joulin et al., 2016) for text, VGGish (Hershey et al., 2017) and Beats (Chen et al., 2022) for acoustic, and ResNet (He et al., 2016) and CLIP (Radford et al., 2021) for visual.

Implementation Details. All experiments were conducted on an NVIDIA RTX A5000 GPU with 24GB of memory. For CARER, ImageNet, and ESC-50, the detailed hyperparameter settings are as follows: number of attention heads {4, 1, 1}, number of attention layers {1, 3, 3}, batch size 64, and learning rate 1e-3. For CMU-MOSI and CMU-MOSEI, the detailed hyperparameter settings are as follows: feature dimension {300, 228}, number of fusion layers {6, 3}, regulation coefficient 0.1, batch size 128, and learning rate 1e-3.

5 Experimental Results

5.1 Mixed Incomplete Data Scenarios

Given the large volume and complexity of video data, partial frame loss frequently occurs to save

Dataset	FG ratio	CG	T^2DR		CubeMLP		MMIN		GMC		RedCore	
	(r_t, r_a, r_v)	ratio	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CMU-MOSI	(0.5, 0.5, 0.5)	0	0.5875	**0.5549**	0.5700	0.5526	0.5816	0.5273	0.5743	0.4783	**0.5977**	0.5275
	(0.1, 0.5, 0.9)	0	0.6735	0.656	0.6866	0.6720	0.6822	0.6622	0.5685	0.4371	**0.6968**	**0.6790**
	(0.1, 0.1, 0.1)	0.1	**0.6866**	**0.6697**	0.6822	0.6672	0.6706	0.6560	0.6706	0.6533	0.6633	0.6494
	(0.5, 0.5, 0.5)	0.5	**0.5758**	**0.5326**	0.5510	0.5266	0.5685	0.5135	0.5685	0.4792	0.5000	0.4817
	(0.9, 0.9, 0.9)	0.9	**0.5539**	**0.4792**	0.4883	0.4686	0.5044	0.4640	0.5510	0.3965	0.4650	0.4555
	(0.1, 0.5, 0.9)	0.5	0.6458	0.6308	**0.6545**	**0.6385**	0.6385	0.6143	0.5918	0.4911	0.5816	0.5705
	(0, 0, 0)	0.1	**0.7259**	**0.7103**	0.6939	0.6791	0.6983	0.6833	0.7099	0.6949	0.6895	0.6752
	(0, 0, 0)	0.5	**0.7041**	**0.6886**	0.6691	0.6542	0.6647	0.6512	0.6924	0.6771	0.5685	0.5575
	(0, 0, 0)	0.9	**0.6603**	**0.6461**	0.6356	0.6146	0.6487	0.6355	0.6516	0.6206	0.4985	0.4884
	(0, 0, 0)	0	**0.7230**	**0.7075**	0.6968	0.6817	0.7055	0.6904	0.7114	0.6963	0.7026	0.6880
CMU-MOSEI	(0.5, 0.5, 0.5)	0	**0.5192**	**0.4978**	0.5166	0.3994	0.4583	0.4167	0.4922	0.4278	0.5132	0.3905
	(0.1, 0.5, 0.9)	0	**0.5941**	**0.575**	0.5793	0.4940	0.5662	0.5520	0.5510	0.5088	0.5782	0.4943
	(0.1, 0.1, 0.1)	0.1	**0.6158**	**0.5971**	0.5765	0.4978	0.5982	0.5609	0.5973	0.5720	0.5626	0.4821
	(0.5, 0.5, 0.5)	0.5	**0.5179**	**0.4892**	0.5158	0.4077	0.4595	0.4185	0.4610	0.4247	0.4671	0.3803
	(0.9, 0.9, 0.9)	0.9	0.4407	**0.4247**	**0.4754**	0.3441	0.3994	0.3828	0.4550	0.3726	0.4226	0.3610
	(0.1, 0.5, 0.9)	0.5	**0.5737**	**0.551**	0.5624	0.4734	0.5327	0.5181	0.5310	0.4981	0.4967	0.4231
	(0, 0, 0)	0.1	**0.6469**	**0.6292**	0.5909	0.5129	0.6255	0.5892	0.6177	0.5892	0.5823	0.5019
	(0, 0, 0)	0.5	**0.6272**	**0.6114**	0.5585	0.4892	0.5911	0.5553	0.6033	0.5864	0.5123	0.4398
	(0, 0, 0)	0.9	**0.6014**	**0.5855**	0.5201	0.4590	0.5759	0.5415	0.5722	0.5628	0.4376	0.3779
	(0, 0, 0)	0	**0.6516**	**0.6345**	0.6029	0.5224	0.6287	0.5931	0.6270	0.5968	0.6040	0.5221

Table 1: Performance analysis of combining fine-grained (FG) and coarse-grained (CG) incomplete data.

bandwidth and computational resources. Thus, we assign the higher missing rate to the video in the imbalanced missing pattern. As shown in Table 1, T^2DR achieves a new state-of-the-art performance across most missing scenarios and even the complete scenario. Specifically, in MOSI, T^2DR improves the Acc by an average of 2.85% and the F1 score by 2.99% over the best-performing MMIN. In MOSEI, T^2DR enhances the Acc by an average of 5.14% and the F1 score by 9.5% over the best-performing GMC. The above results demonstrate that it is effective to allocate more attention to the more crucial data when dealing with incomplete multimodal data. Notably, all comparative algorithms exhibit a significant difference between F1 and Acc scores in MOSEI. To figure out this matter, we perform detailed experiments in Appendix D.2 and find out that it is because the larger MOSEI has a more balanced distribution of three classes, which amplifies the impact of category performance imbalances on the weighted F1 score. Fortunately, in CAS, the secondary supervision provided by the CE loss function effectively enhances focus on each sample's classification, achieving relatively balanced performance across different classes.

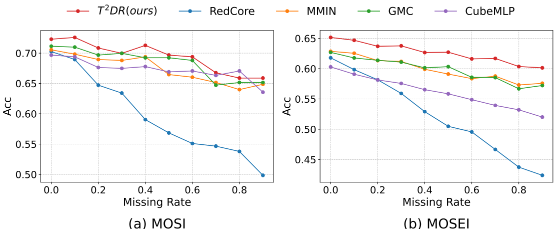


Figure 3: Coarse-grained effectiveness evaluation.

5.2 Coarse-grained Incomplete Data Scenarios

To sufficiently demonstrate T^2DR's effectiveness in coarse-grained incomplete data scenarios, we expand the comparison in Table 1 by including additional levels of modality missing rates.

The comparison of accuracy results is shown in Figure 3, which reveals that *IEDR* continues to exhibit superior performance compared to other competitive methods. The specifics are detailed below: (i) While maintaining the absolute superiority of MOSEI and the overall advantage of MOSI, T^2DR does not experience significant drops at any specific missing rate, demonstrating exceptional robustness across different degrees of deficiency. (ii) Even as the missing rate rises, reducing the amount of learnable information, occasional slight

Table 2: Performance analysis of fine-grained imputation in various modalities.

Dataset	Method	Deficiency Rate r_{fg}							Average Improvement Rate
		0	0.1	0.3	0.5	0.7	0.9	AVG	
CARER	TextCNN (Gong and Ji, 2018)	0.8420	0.8100	0.7200	0.6195	0.5370	0.3990	0.6546	+4.43%
	+ IADR	**0.8745**	**0.8500**	**0.7515**	**0.6535**	**0.5490**	**0.4230**	**0.6836**	
	FastText (Joulin et al., 2016)	0.6660	0.6695	0.6405	0.6085	**0.5965**	0.5715	0.6254	+2.51%
	+ IADR	**0.7005**	**0.6900**	**0.6565**	**0.6230**	0.5950	**0.5815**	**0.6411**	
ImageNet	ResNet (He et al., 2016)	0.6690	0.6515	0.6101	0.5394	0.4188	0.1943	0.5139	+3.25%
	+IADR	**0.6869**	**0.6675**	**0.6280**	**0.5566**	**0.4372**	**0.2076**	**0.5306**	
	CLIP (Radford et al., 2021)	**0.6180**	**0.5776**	0.3658	0.1251	0.0157	0.0018	0.284	+38.59%
	+ IADR	0.6167	0.5761	**0.4994**	**0.3874**	**0.2396**	**0.0424**	**0.3936**	
ESC-50	VGGish (Hershey et al., 2017)	0.5000	0.5950	0.6250	0.5600	0.2750	0.0500	0.4342	+11.88%
	+ IADR	**0.6150**	**0.6150**	**0.6300**	**0.6800**	**0.3100**	**0.0650**	**0.4858**	
	Beats (Chen et al., 2022)	0.7500	0.7650	0.7250	0.7300	**0.7000**	**0.5300**	0.7000	+3.1%
	+ IADR	**0.8150**	**0.8050**	**0.7550**	**0.7450**	0.6850	0.5250	**0.7217**	

increases in both benchmarks suggest that T^{2}DR successfully learns valuable information from other modalities, thereby alleviating the loss of information due to reduced data. (iii) On the larger MOSEI, with the optimal performance achieved, the smallest accuracy reduction (5.02%) from the highest missing rate to the lowest is realized, highlighting T^{2}DR's superior performance.

5.3 Fine-grained Incomplete Data Scenarios

To further evaluate T^{2}DR's applicability to single-modal data, we conduct intra-modal incomplete data experiments on text, visual, and audio modalities. The *IADR* module is tested on datasets from all three modalities, with two mainstream models validated for each modality. Furthermore, we also compare its performance with masked multi-head attention in Appendix D.1, demonstrating the necessity of retaining global capabilities.

As shown in Table 2, the experimental results indicate that: (i) Intra-Attn shows excellent compatibility across six different methods for three modalities in fine-grained missing scenarios, confirming its effectiveness in addressing intra-modal missing data for various modalities and methods. (ii) The average performance of organized data such as images and audio is slightly better than that of unorganized text, suggesting that the mask with explicit missing information can effectively provide more targeted data supervision. (iii) Notably, on the ImageNet dataset, the CLIP+IADR method achieves an average improvement rate of 38.59% over the original CLIP, suggesting that the IADR module can significantly enhance the performance of models sensitive to incomplete data.

Table 3: The ablation study on MOSEI with six scenarios: coarse-grained deficiency rates r_{cg} are c_0, c_1, c_2, and c_3 respectively, while the fine-grained deficiency rates r_{fg} are $f_0 = (c_0, c_0, c_0)$ and $f_1 = (c_1, c_2, c_3)$, where c_0 = 0, c_1 = 0.1, c_2 = 0.5, and c_3 = 0.9.

Model	**Testing Conditions** $(r_{fg}; r_{cg})$					
	$(f_1; c_0)$	$(f_1; c_1)$	$(f_1; c_2)$	$(f_1; c_3)$	$(f_0; c_2)$	$(f_0; c_0)$
T^{2}DR	**0.5941**	**0.5907**	**0.5737**	**0.5621**	**0.6272**	**0.6516**
w/o*IADR*	0.5842	0.5812	0.5641	0.5458	0.6162	0.6428
w/o *CAS*	0.583	0.5812	0.5671	0.5566	0.6137	0.6407
w/o *SFP*	0.5933	0.5857	0.5377	0.5003	0.5785	0.6499
w/o*IEDR*	0.4449	0.4346	0.4288	0.4248	0.5119	0.5681

5.4 Ablation Studies

To verify the necessity of different components, we conduct ablation studies on MOSEI under the above three incomplete data scenarios. As shown in Table 3, we can find below: (i) Intra-Modal Deficiency-Resistant (IADR) Removal: The performance decline under fine-grained deficiency indicates that IADR's balance between the missing parts' effect and model's global capacity effectively resists the noise impact from incomplete data. (ii) Share Feature Prediction (SFP) Removal: The larger performance decline with inter-modal deficiency reveals that SFP's prediction on the missing modality is crucial to avoid the zero-information issue for course-grained incomplete data. (iii) Capability-Aware Scorer (CAS) Removal: The relatively balanced performance decline across different conditions suggests that the dynamic modality weights perceived by CAS are essential for Inter-Attn to integrate modality features in any situation. (iv) Inter-Modal Deficiency-Resistant (IEDR) Removal: The significant performance decrease reveals that the semantics of the encoded data are insufficient to directly perform downstream tasks.

6 Conclusion

In this paper, we propose T^{2}DR, a comprehensive framework for addressing both intra-modality and inter-modality incomplete data. To address various granularities of deficiency, we introduce two weight allocation mechanisms, Intra-Attn and Inter-Attn, to provide stronger supervision of higher-quality data. Additionally, for coarse-grained deficiencies, we present shared feature prediction (SFP) for missing modality imputation, and capability-aware scorer (CAS) to dynamically perceive the optimal weight for each modality, enabling Inter-Attn-based adaptive supervision. Our experiments on both unimodal and multimodal datasets have demonstrated the remarkable robustness and adaptability of T^{2}DR under incomplete conditions.

7 Limitations

We have conducted extensive experiments across fine-grained, coarse-grained, and mixed deficiencies, demonstrating the remarkable robustness and adaptability of our proposal. However, it still encounters limitations when handling imbalanced missing data in the smaller MOSI dataset. This issue arises mainly from two factors: (i) With imbalanced missing data across different modalities, T^{2}DR tends to overly rely on data-rich modalities on the smaller datasets, which weakens the inter-modal interactive learning. (ii) T^{2}DR's employment of Inter-Attn to assign greater weight to more effective modalities further amplifies the problem of neglecting less data-abundant modalities. This limitation reveals the need for strategies that enforce inter-modal interaction, thereby facilitating the learning of more comprehensive cross-modal complementary information.

Building on the foundation of providing greater supervision to effective data, our future work will focus on enhancing inter-modal interaction to fully capture cross-modal complementary information. By enabling the above, our approach will better adapt to multimodal imbalanced missing data in small datasets, achieving greater robustness in incomplete multimodal learning.

8 Acknowledgments

This work was supported by the "Pioneer" R&D Program of Zhejiang (No.2024C01021), Zhejiang Province "Jianbing" Key R&D Project of China (No.2025C01010), and the Major Research Program of Zhejiang Provincial Natural Science Foundation (No. LD24F020015). The authors are supported by Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security.

References

Iman Ahmadianfar, Ali Asghar Heidari, Saeed Noshadian, et al. 2022. Info: An efficient optimization algorithm based on weighted mean of vectors. *Expert Systems with Applications*, 195:116516.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *WACV*, pages 1–10.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, et al. 2021. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *ICCV*, pages 4015–4025.

Sanyuan Chen, Yu Wu, Chengyi Wang, et al. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Yen-Chun Chen, Linjie Li, Licheng Yu, et al. 2020. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120.

Jia Deng, Wei Dong, Richard Socher, et al. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *MM*, pages 1459–1462.

Ming Fang, Siyu Peng, Yujia Liang, et al. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.

Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. 2021. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, pages 15506–15515.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Linyuan Gong and Ruyi Ji. 2018. What does a textcnn learn? *arXiv preprint arXiv:1801.06287*.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1726–1736. ACL.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on CVPR*, pages 770–778.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, et al. 2017. Cnn architectures for large-scale audio classification. In *icassp*, pages 131–135.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kenneth Lau, Jonas Adler, and Jens Sjölund. 2019. A unified representation network for segmentation with missing modalities. *arXiv preprint arXiv:1908.06683*.

Gwen Legate, Lucas Caccia, and Eugene Belilovsky. 2023. Re-weighted softmax cross-entropy to control forgetting in federated learning. In *Conference on Lifelong Learning Agents*, pages 764–780.

Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *CVPR*, pages 10758–10768.

Chengliang Liu, Jie Wen, Zhihao Wu, Xiaoling Luo, Chao Huang, and Yong Xu. 2023. Information recovery-driven deep incomplete multiview clustering network. *IEEE Transactions on Neural Networks and Learning Systems*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

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

Xianping Ma, Xiaokang Zhang, Man-On Pun, and Ming Liu. 2024. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*.

Arsha Nagrani, Shan Yang, Anurag Arnab, et al. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213.

Shah Nawaz, Muhammad Saad Saeed, Muhammad Zaigham Zaheer, et al. 2024. Gazelle: A multimodal learning system robust to missing modalities.

Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia*, 4(2):1883.

Petra Poklukar, Miguel Vasco, Hang Yin, et al. 2022a. Geometric multimodal contrastive representation learning. In *ICML*, pages 17782–17800.

Petra Poklukar, Miguel Vasco, Hang Yin, et al. 2022b. Geometric multimodal contrastive representation learning. In *ICML*, pages 17782–17800.

Jiahui Qu, Yuanbo Yang, Wenqian Dong, and Yufei Yang. 2024. Lds2ae: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities. In *AAAI*, volume 38, pages 14731–14739.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Mohammad Reza Rezaei-Dastjerdehei, Amirmohammad Mijani, and Emad Fatemizadeh. 2020. Addressing imbalance in multi-label classification using weighted cross entropy loss function. In *ICBME*, pages 333–338.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, et al. 2018. Carer: Contextualized affect representations for emotion recognition. In *EMNLP*, pages 3687–3697.

Gwenaelle Cunha Sergio and Minho Lee. 2021. Stacked debert: All attention in incomplete data for text classification. *Neural Networks*, 136:87–96.

Chao Shang, Aaron Palmer, Jiangwen Sun, Ko-Shin Chen, Jin Lu, and Jinbo Bi. 2017. Vigan: Missing view imputation with generative adversarial networks. In *Big Data*, pages 766–775.

Siddharth Srivastava and Gaurav Sharma. 2024. Omnivec2-a novel transformer based network for large scale multimodal and multitask learning. In *CVPR*, pages 27412–27424.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *MM*, pages 3722–3729.

Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. 2024. Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *AAAI*, volume 38, pages 15173–15182.

Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021. Multimodal cross-and self-attention network for speech emotion recognition. In *ICASSP*, pages 4275–4279.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Huayi Tang and Yong Liu. 2022. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *ICML*, pages 21090–21110.

Tim Van Erven and Peter Harremos. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinhang Wan, Xinwang Liu, Jiyuan Liu, et al. 2023. Auto-weighted multi-view clustering for large-scale data. In *AAAI*, volume 37, pages 10078–10086.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Multimodal learning with missing modality via shared-specific feature modelling. pages 15878–15887.

Shicai Wei, Chunbo Luo, and Yang Luo. 2023a. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *CVPR*, pages 20039–20049.

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. 2023b. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st Annual Meeting of the ACL*, pages 5240–5252.

Xiaohan Xing, Zhen Chen, Meilu Zhu, et al. 2022. Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 636–646.

Cai Xu, Ziyu Guan, Wei Zhao, Hongchang Wu, Yunfei Niu, and Beilei Ling. 2019. Adversarial incomplete multi-view clustering. In *IJCAI*, volume 7, pages 3933–3939.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. 2024. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *AAAI*, volume 38, pages 16416–16424.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2023. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539.

Ziqi Yuan, Baozheng Zhang, Hua Xu, and Kai Gao. 2024. Meta noise adaption framework for multimodal sentiment analysis with feature noise. *IEEE Transactions on Multimedia*, 26:7265–7277.

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

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, et al. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, pages 2236–2246.

Jingyi Zhang, Zhen Wei, Ionut Cosmin Duta, Fumin Shen, Li Liu, Fan Zhu, Xing Xu, Ling Shao, and Heng Tao Shen. 2019. Generative reconstructive hashing for incomplete video analysis. In *MM*, pages 845–854.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021a. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021b. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *ACL*, pages 2608–2618.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021c. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *ACL*, pages 2608–2618.

Heqing Zou, Yuke Si, Chen Chen, et al. 2022. Speech emotion recognition with co-attention based multi-level acoustic information. In *ICASSP*, pages 7367–7371.

A Supplementary for Methods

A.1 Sparsity-Induced Attention Collapse

In this section, we explore how masked multi-head attention performs under varying degrees of data incompleteness. Intuitively, completely concentrating on non-missing segments may weaken the model's overall ability to capture global dependencies. To validate this hypothesis, we introduce the index "attention distance", which quantifies how far each token can effectively attend to others (Dosovitskiy et al., 2020), providing an intuitive measure of the model's global capacity. Initially, we provide the theoretical proof for this hypothesis, analyzing how attention distance – the model's global capacity changes with varying degrees of data incompleteness. Since attention distance was first proposed in ViT without a generalized computation pipeline for other modalities, we focus our certification on the visual modality. To compute the attention distance, we first compute the spatial arrangement of image patches. Given an image divided into N non-overlapping patches, arranged in a $L \times L$ grid where $L = \sqrt{N}$, the (i, j)-th patch is positioned at:

$$(x_i, y_i) = \left(\left\lfloor \frac{i}{L} \right\rfloor, i \mod L \right). \tag{15}$$

The Euclidean distance between patches i and j is then computed as:

$$D_{i,j} = p \cdot \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \tag{16}$$

where p denotes the patch size. This results in a distance matrix $D \in \mathbb{R}^{N \times N}$, capturing the pairwise spatial distances between all patches, while the attention score A dictates their interaction capacity; thus integrate both as follows:

$$M = A \odot D, \tag{17}$$

where $\odot$ denotes the element-wise product. To obtain the average distance attended by each token, we need to sum along the last axis:

$$M' = \sum_{j=1}^{N} M_{i,j}, \quad \forall i \in \{1, \ldots, N\}. \tag{18}$$

Finally, the mean attention distance (MAD) is computed by averaging over all tokens:

$$\text{MAD} = \frac{1}{N} \sum_{i=1}^{N} M'_i. \tag{19}$$

The above defines the complete computation process of attention distance. However, as the missing rate increases, the attention score A will change. The detailed process is as follows.

$$A = \text{softmax}\left(\frac{QK^T + M}{\sqrt{d}} \right), \tag{20}$$

$$M = \begin{cases} -\infty, & \text{if the patch is missing,} \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

As the missing rate increases, a growing number of patches become unavailable, causing more attention scores to drop to zero, as demonstrated in Equations (20) and (21). This progressive decline in attention scores consequently causes a significant reduction in M in Equation (17), ultimately resulting in a substantial decrease in the final attention distance.

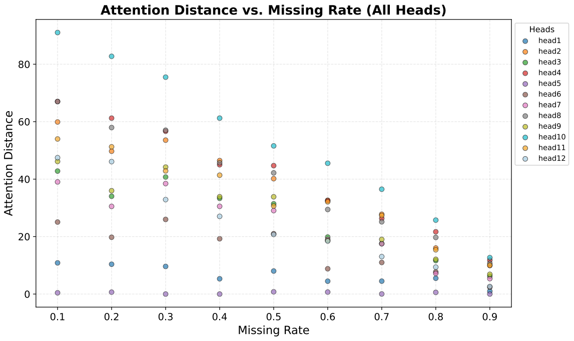


Figure 4: Variation of Attention Distance Across All Attention Heads Under Different Missing Rates.

This theoretical finding was further validated through experimental evaluation. We systematically assessed the attention distance of vit-base-patch16-224 on 1,000 images from the ImageNet test set under varying missing rates from 0 to 0.9. As shown in Figure 4, the attention distance of each head exhibits a significant decreasing trend as the missing rate increases. This observation further validates our hypothesis.

Moreover, we performed a comparative analysis of the head-averaged attention distance between Intra-Attn and the original method, as depicted in Figure 6. The results demonstrate that Intra-Attn achieves a notably higher average attention distance, providing compelling evidence for its effectiveness in enhancing the model's global reasoning capability.

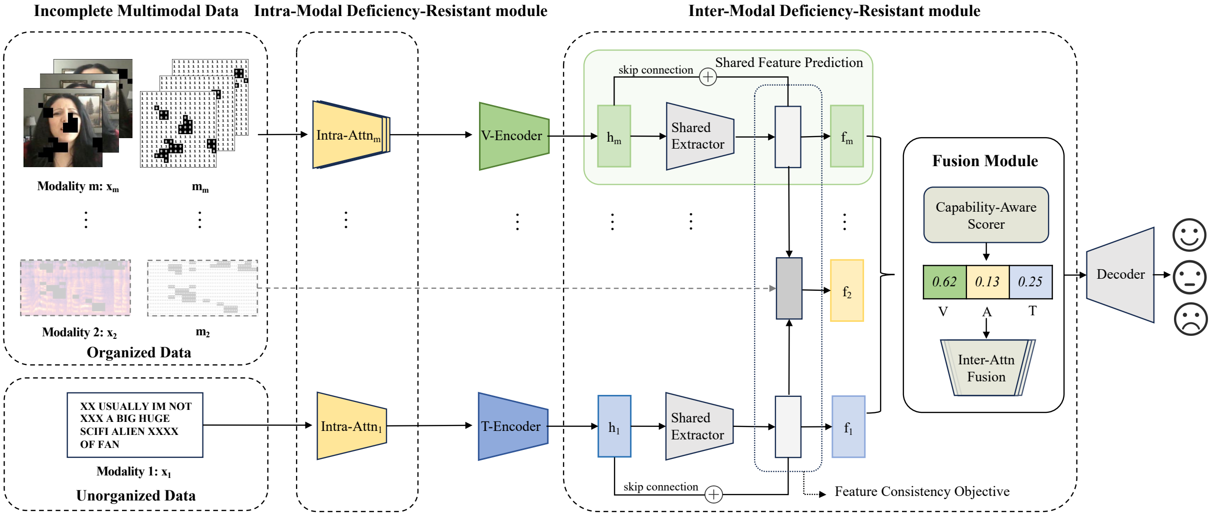


Figure 5: **An overview of T^{2}DR in the context of inter-modality incomplete data.**

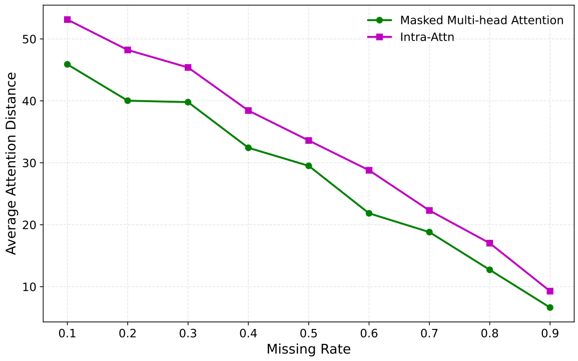


Figure 6: Comparison of Average Attention Distance Between Our Method and the Baseline.

A.2 Intra-Modal Deficiency-Resistant module

In this section, we present a detailed description of the processing of Intra-Modal Deficiency-Resistant (IADR) across text, visual, and acoustic modalities. For unorganized text, token-mapped embeddings are fed into Intra-Attn before encoding to filter out noisy data. For organized visual and acoustic, the pixelated images (or spectrograms) and their corresponding masks are divided into patches and undergo a window-shifting operation, after which they are fed into Intra-Attn to highlight available information without disrupting model global capability. In summary, IADR incorporates Intra-Attn to perform selective information filtering on raw digital modality inputs prior to feature encoding, which effectively addresses the challenges of data incompleteness and noise contamination that typically degrade model performance in incomplete multi-modal learning.

A.3 Inter-Modal Deficiency-Resistant module

In this section, we provide a detailed description of the Inter-Modal Deficiency-Resistant module (IEDR), including the pseudocode 1 of IEDR and the T^{2}DR framework in Figure 5 for inter-modality incomplete data scenarios.

Algorithm 1 Inter-Modal Deficiency-Resistant

```
Input:
    Unified features of various modalities:
      $\mathcal{H} = \{h_i^{(k)} \in \mathbb{R}^d\}_{i,k=1}^{m,n}$
    Entire modality mask of various modalities:
      $\mathcal{M} = \{m_i^{(k)} \in \{\text{True}, \text{False}\}\}_{i,k=1}^{m,n}$
Output:
    Fused feature across various modalities:
      $\mathcal{F} = \{f^{(k)} \in \mathbb{R}^{m \times d}\}_{k=1}^{n}$
/* Script symboys $i$ and $k$ mean index of modality and instance. If $m_i$ is True, it indicates the corresponding modality is present; if False, it is missing. */
for $k = 1$ to $n$ do
    for $i = 1$ to $m$ do
        if $m_i$ then
            $f_i \leftarrow \text{SFP}(h_i)$
        else
            $f_i \leftarrow \text{mean}(S(h_i))$ for matching $m_i$ is True
        end if
    end for
    $weights \leftarrow \text{CAS}(f_1 \oplus f_2 \oplus \cdots \oplus f_m)$
    $f^{(k)} \leftarrow \text{Inter-Attn}\{(f_1 \oplus f_2 \oplus \cdots \oplus f_m), weights\}$
end for
$k = 1, 2, \ldots, n; i = 1, 2, \ldots, m$
```

B Supplementary for Baselines

We implemented two types of baselines, covering intra- and inter-modal incomplete data. The details of each baseline are listed below:

Multimodal Models. *CubeMLP* (Sun et al., 2022) integrates modality features into a unified tensor and applies sequential, modality, and channel mixing through independent MLP units, effectively enhancing feature interaction and reducing computational complexity. *MMIN* (Zhao et al., 2021c) uses a Cascade Residual Autoencoder and Cycle Consistency Learning to predict missing modality representations from available ones, enhancing robustness and adaptability under uncertain missing-modality conditions. *GMC* (Poklukar et al., 2022b) utilizes modality-specific base encoders and a shared projection head to align representations of complete and modality-specific observations, employing a novel multimodal contrastive loss to address the heterogeneity gap and ensure robustness to missing modality information. *RedCore* (Sun et al., 2024) leverages a relative advantage-aware cross-modal representation learning framework based on variational information bottleneck (VIB), coupled with a bi-level optimization problem for adaptive supervision regulation, ensuring robustness and data efficiency.

Unimodal Models. *TextCNN* (Gong and Ji, 2018) employs convolutional neural networks to capture local dependencies and extract semantic features from text, utilizing multiple filter sizes to identify important n-grams, and offering a robust architecture for sentence classification tasks. *FastText* (Joulin et al., 2016) innovatively merges a linear classifier, low-rank approximation, and hierarchical softmax to achieve efficient, scalable text classification, improving performance by using bag-of-n-grams as features to capture local word order. *ResNet* (He et al., 2016) employs deep residual learning through shortcut connections to effectively train very deep networks, enabling improved optimization and higher accuracy by reformulating layers to learn residual functions instead of direct mappings. *CLIP* (Radford et al., 2021) leverages contrastive pre-training on a large dataset of image-text pairs to enable zero-shot image classification by aligning visual and textual representations within a shared multimodal embedding space. *VGGish* (Hershey et al., 2017) is a pre-trained audio encoder based on the extensive AudioSet, designed to generate a 128-dimensional feature embedding for enhanced performance across various audio recognition tasks. *Beats* (Chen et al., 2022) combines an acoustic tokenizer and an audio SSL model to convert continuous audio signals into semantic-rich discrete labels, enhancing the model's audio understanding capabilities through mask and label prediction pre-training.

C Supplementary for Datasets

Multimodal Datasets. The CMU-MOSI dataset comprises 2,199 annotated video segments with sentiment intensity labels, whereas CMU-MOSEI extends this scale with 23,453 video clips containing both sentiment and emotion annotations. Both datasets maintain predefined training, validation, and test splits as specified in their official documentation.

Unimodal Datasets. ImageNet (1.43M images), CARER (20k emotional text), and ESC-50 (2k environmental audio clips) all adhere to standardized partitioning protocols established by their respective curators.

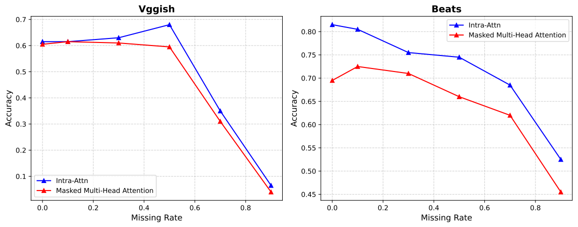


Figure 7: Comparative Analysis of Improved Intra-Attn and Masked Multi-Head Attention in Experimental Results.

D Supplementary for Experiments

D.1 Supplementary for IADR Experiments

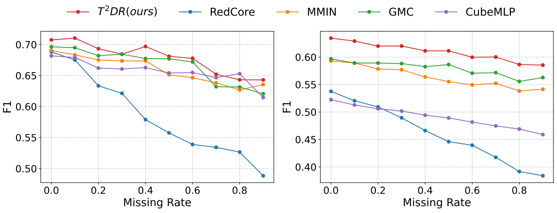
As discussed in Appendix A.1 , in scenarios with missing data, the direct application of conventional attention mechanisms that exclusively rely on available data would inevitably lead to significant degradation in the model's global capability. To address this limitation, we propose Intra-Attn, which introduces a compensation term for missing data to balance noise interference and global performance degradation. (1) In visual tasks, our method achieves consistent improvements across baseline architectures (average +1.1% on CLIP and +0.1% on ResNet). (2) In audio tasks, Intra-Attn demonstrates substantially enhanced robustness to

Table 4: Performance analysis of Intra-Modal Deficiency-Resistant module in various modalities.

Dataset	Method	Deficiency Rate r_{fg}							Average Improvement Rate
		0	0.1	0.3	0.5	0.7	0.9	AVG	
ImageNet	ResNet+KNN	0.7273	0.7087	0.6427	0.5118	0.3657	0.1212	0.5129	-0.19%
	CLIP+KNN	0.6181	0.5679	0.4512	0.2923	0.1382	0.0019	0.3449	+21.44%
ESC-50	VGGish+KNN	0.4950	0.5400	0.5800	0.5600	0.2050	0.0450	0.4042	-6.91%
	Beats+KNN	0.7600	0.7700	0.7400	0.7300	0.6700	0.5100	0.6967	-0.47%

missing spectral features, as illustrated in Figure 7. These quantitative results and comparative analyses collectively demonstrate that Intra-Attn not only effectively mitigates the impact of missing data within a modality but also preserves the model's overall capability.

Based on the experiments with intra-modality incomplete data, we also compare these results with the simple imputation method K-Nearest Neighbors (KNN) (Peterson, 2009). It fills in missing data by averaging the values of the nearest neighbors, ensuring the integrity of the dataset. Since the imputation process depends on mask information, which is not feasible for unstructured text, we focus our experiments exclusively on the visual and acoustic modalities. As shown in Table 4, which supplements the results of the section 5.3, we find that KNN only works for the CLIP model, which is extremely sensitive to incomplete data (with an accuracy drop of 0.4929 from $r_{fg} = 0$ to $r_{fg} = 0.5)$. However, its average improvement rate of 21.54% is still significantly lower than the 38.59% achieved by IADR. KNN even undermines the performance of the original model in most cases, likely due to the increased filling errors in high-dimensional or imbalanced data distributions. Hence, it is quite remarkable for IADR to exhibit such strong robustness in handling intra-modality incomplete data across six classical methods in three modalities.



(a) MOSI

(b) MOSEI

Figure 8: Coarse-grained effectiveness evaluation on F1 score.

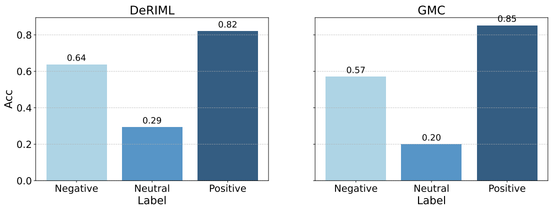


Figure 9: Label-specific Accuracy Comparison on MOSEI.

D.2 Supplementary for IEDR Experiments

Following the experiments with inter-modality incomplete data, we conduct an additional evaluation of the F1 metric shown in Figure 8. We find that, compared to MOSI, T^{2}DR demonstrates a more dominant performance in MOSEI, which is attributed to two key factors: (i) T^{2}DR, as an attention-based method, is better suited for large-scale datasets (Dosovitskiy et al., 2020). (ii) Our approach achieves a more balanced performance across various categories in MOSEI. Specifically, with the significant increase in neutral samples in MOSEI (from 4.41% to 21.87%), however, neutral sentiment is hard to classify due to the lack of strong sentiment indicators and the difficulty in detecting weakly associated cues (Gandhi et al., 2023). Therefore, achieving a higher F1 score on MOSEI requires more balanced performance across all three categories, as weak classification of neutral sentiment will be distinctly amplified in F1. To validate the above idea, we compare T^{2}DR with subsequent GMC and notice that when the dataset is complete, the drop from F1 to accuracy is 1.71% and 3.02% respectively. This large disparity drives us to conduct a more thorough analysis. We compute the prediction result for each label shown in Figure 9, which reveals that T^{2}DR presents smaller performance variation across various classes, resulting in a higher F1 despite having consistent accuracy.