DEAR: Distributional Error-Aware Reliability for Robust Multimodal Sentiment Analysis with Missing Modalities

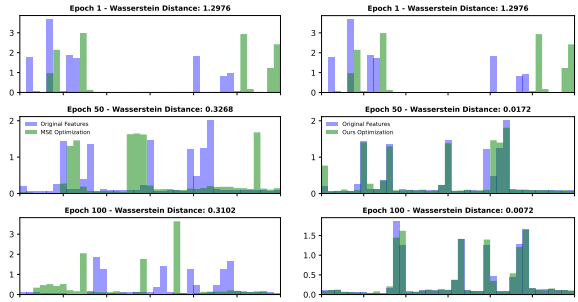
Shihao Zou[1,2], **Wei Wei**[1,2]* **and Yongshuo Zhang**[1,2]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology,
[2]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL).
{sh_zou, weiw, zyshuo}@hust.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) often suffers from performance degradation due to missing modalities in practical applications. Existing methods typically focus on feature completion but neglect semantic shifts caused by distribution gaps and decision risks under high uncertainty. In this paper, we propose a **D**istributional **E**rror-**A**ware **R**eliability (DEAR) estimation framework for robust MSA. Specifically, we design a Hierarchical Distribution-Constrained Reconstruction (HDCR) module to mitigate semantic shifts by explicitly aligning reconstructed features with the original distributional manifold. Meanwhile, a reliability evaluation module (SURE) is introduced to quantitatively measure reconstruction fidelity. By perceiving inherent uncertainty, SURE provides a reliability-driven gating mechanism for the Synergistic-Robust Dual-Stream (SR-DS) architecture. This mechanism enables the model to dynamically adjust contribution weights: strengthening cross-modal synergistic effects when data fidelity is high, while shifting focus toward robust paths under high-risk missingness to safeguard performance. Extensive experiments on MOSI, MOSEI, and SIMS datasets validate the effectiveness and decision reliability of DEAR.

1 Introduction

Multimodal sentiment analysis (MSA) infers human affect by integrating heterogeneous signals such as text, vision, and audio, and is fundamental to empathetic human computer interaction. However, cross-modal information is not perfectly complementary, making robust fusion under complex consistency and discrepancy a core challenge (Ramesh et al., 2021; Huang et al., 2021; Ding et al., 2024). In practice, modalities are often missing or degraded due to sensor failures, bandwidth constraints, or privacy protection. This full-modality training, but partial-modality deployment gap can cause drastic performance drops, limiting the reliability of MSA systems.



(a) Current Reconstruct Loss (b) Our Reconstruct Loss

Figure 1: Visualization of feature distributions obtained by calculating the Wasserstein distance for different training stages in the MOSI dataset.

Recent methods aim to enhance joint representation learning or promote cross-modal knowledge transfer, for example, HRLF (Li et al., 2024b) with adversarially regularized hierarchical mutual information, and CorrKD (Li et al., 2024c) via contrastive distillation. Despite progress, two limitations remain. First, many reconstruction branches overly depend on point-to-point losses (e.g., MSE), which may match feature means but distort higher-order statistics, pushing reconstructed features away from the original distributional support and leading to distributional drift (Fig. 1(a)). Second, fusion is often governed by sample-agnostic fixed rules, even though missing-induced shifts vary across samples and modalities: some samples retain key semantics, while others undergo severe manifold collapse. Overlooking such heterogeneous shifts makes models brittle under diverse missing patterns.

Revisiting these bottlenecks from a domain adaptation perspective, we formulate the missing-modality problem as a shift between domains. Specifically, we treat complete modality inputs as

*Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 30344–30361
July 2-7, 2026

the source domain, and the representation space induced by missing inputs as the target domain, attributing performance degradation to representation shift from source to target. Building on this view, we propose a Distributional Error-Aware Reliability (DEAR) estimation framework, which explicitly narrows the distribution gap between the two domains and makes instance-wise decisions according to sample-level transfer difficulty. Concretely, we first design a Hierarchical Distribution-Constrained Reconstruction (HDCR) module to anchor the statistical structure of source-domain features in the representation space. By introducing distribution-level constraints (Fig. 1(b)), HDCR pulls representations of missing views back toward the source manifold, effectively reducing inter-domain discrepancy and preserving the discriminative power of the completed features.

On top of alleviating distribution shift, we further model sample-specific transfer difficulty via a Transfer Difficulty Estimator (TDE). This module directly senses, from missing inputs, how far each modality deviates from the source domain, and maps the estimated shift to fusion weights through a Monotonic Weighting Network (MWN), yielding an interpretable fusion rule: larger shift , lower reliability ,smaller weight. DEAR then introduces a Synergistic-Robust Dual-Stream prediction mechanism: the main stream captures deep cross-modal interactions, while the auxiliary stream performs robust, conservative aggregation guided by MWN-derived reliability weights. Finally, a learnable gating mechanism adaptively balances the two streams, enabling stable generalization under varying degrees of modality absence. Both empirical results and theoretical analysis demonstrate that modeling transfer difficulty at the instance level substantially reduces target-domain risk, providing a more robust solution for multimodal fusion in missing-modality scenarios.

In summary, our main contributions are as follows:

- We propose a HDCR mechanism that introduces distribution consistency constraints to explicitly confine reconstructed features to the original full-modality manifold, thereby effectively eliminating the semantic-space shift caused by missing modalities.
- We design a SURE module to estimate reconstruction difficulty and, together with the SR-DS module, enable seamless arbitration from aggressive cross-modal synergy to conservative uni-modal fallback when high-uncertainty scenarios is detected, substantially improving prediction robustness.
- Extensive experiments on multiple datasets show that DEAR consistently outperforms the SOTA baselines across various metrics, and qualitative visualizations further verify its decision stability under different missing-modality scenarios.

2 Related Works

2.1 Multimodal Sentiment Analysis

Early sentiment analysis primarily relied on textual data (Yin and Zhong, 2024; Chen et al., 2024; He et al., 2024),which often suffered from ambiguity. Multimodal Sentiment Analysis (MSA) addresses this by integrating heterogeneous sources(text, audio, video) to capture complementary emotional cues. Current mainstream research focuses on designing sophisticated fusion mechanisms and interaction paradigms (Tsai et al., 2019; Hazarika et al., 2020; Liang et al., 2020; Rahman et al., 2020; Yu et al., 2021; Han et al., 2021; Lv et al., 2021; Yang et al., 2022; Guo et al., 2022; Sun et al., 2022; Li et al., 2023; Zhang et al., 2023, 2025; Tian et al., 2025) , aiming to better exploit the complementary relationships between modalities and thereby improve overall performance. For example, MulT (Tsai et al., 2019) utilizes directed pairwise cross-modal attention to handle unaligned sequences and long-range dependencies. CubeMLP (Sun et al., 2022) proposes a lightweight fully MLP-based framework that mixes multimodal features along three axes to capture diverse interactions. DMD (Li et al., 2023) enhances modality-specific discriminative power through flexible cross-modal distillation in decoupled representation spaces. Furthermore, SDRS (Zhao et al., 2025) refines sentiment analysis by extracting emotion-specific representations to dynamically shift textual features in the latent label space.

2.2 MSA with Missing Modalities

In MSA tasks, modality missing scenarios better reflect real-world applications but also introduce greater challenges. Recent work has made notable progress in addressing this issue (Yuan et al., 2021; Li et al., 2024a; Lian et al., 2023; Pham et al., 2019; Liu et al., 2024). Some approaches focus on data level recovery; for instance, DiCMoR (Wang et al., 2023c) transfers distributions from available modalities to missing ones to maintain global con-

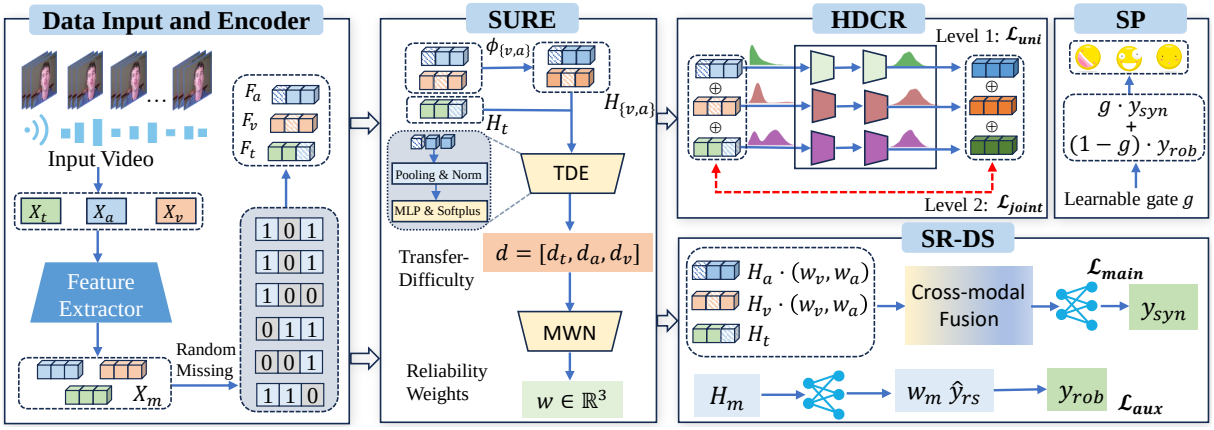


Figure 2: The overall architecture of our proposed model. SURE denotes the *Sample-Level Uncertainty and Reliability Estimator* module, HDCR denotes the *Hierarchical Distribution Constraint Reconstruction* module, SR-DS denotes the *Synergistic-Robust Dual-Stream Fusion* module, and SP denotes the *Sentiment Prediction* module.

sistency. Others prioritize representation robustness, such as NIAT (Yuan et al., 2024), which employs adversarial training to align noisy and original instances, and UMDF (Li et al., 2024a), which learns inherent multimodal representations from consistent distributions. Furthermore, several recent studies have utilized feature decoupling or advanced alignment strategies to enhance robustness. CorrKD (Li et al., 2024c) proposes a correlation-decoupled knowledge distillation framework that transfers cross-sample and category knowledge to reconstruct missing semantics and optimize sentiment decision boundaries. LNLN (Zhang et al., 2024) designates language as the dominant modality and introduces a correction and a language-driven learning module to ensure robustness across various noise scenarios. HRLF (Li et al., 2024b) adopts a hierarchical representation learning framework that reconstructs high-level semantics through cross-modal translation and hierarchical mutual information maximization. P-RMF (Zhu et al., 2025) takes a data-centric approach by mapping unimodal data to Gaussian latent spaces and learning a stable proxy modality representation based on quantified intrinsic uncertainty. DAR (Yang et al., 2025) explicitly addresses modality interplay through a mutual information-based decoupling module and independent reconstruction of common and specific representations.

Despite these advancements, existing methods face two primary limitations: (i) Feature restoration often focuses on numerical completion rather than statistical anchoring, leading to persistent semantic shifts. (ii) Robust architectures typically rely on static fusion logic, failing to protect the decision lower bound in high-uncertainty scenarios. In contrast, our DEAR framework bridges these gaps by enforcing hierarchical distributional constraints for manifold-level repair and introducing a reliability-driven dual-stream architecture to adaptively safeguard prediction stability.

3 Method

In this section, we present DEAR, a distributional error-aware reliability framework for incomplete MSA, shown in Fig. 2. DEAR first repairs collapsed feature manifolds via hierarchical distribution constraints, then quantifies instance-level representation shifts through a self-supervised uncertainty estimator, and finally achieves robust affect prediction using a synergistic robust dual stream fusion strategy.

3.1 Problem Formulation

Given a video clip, we extract complete features for three modalities, denoted as $X_m \in \mathbb{R}^{T_m \times d_m}, m \in \{t, v, a\}$, where T_m and d_m denote the sequence length and feature dimension, respectively. Following prior work LNLN (Zhang et al., 2024), we simulate modality-missing scenarios by applying random missing operations with rates ranging from 0% to 100%. Specifically, we perform random missing in the visual and audio modalities by setting the features to zero, while for the language modality, tokens are replaced with the special [UNK] token to simulate semantic absence. This type of operation corresponds to inter-modal missing, where parts of the data within a modality are randomly removed.

In addition, we also simulate intra-modal missing, where an entire modality is completely dropped from a sample. This allows us to explore the contribution of each individual modality to the MSA task, and evaluate the robustness of our proposed model under various levels and types of modality missing conditions. We denote the incomplete features as F_m. Our goal is to predict sentiment intensity y by effectively sensing and mitigating the uncertainty introduced by F_m.

3.2 Hierarchical Distribution Constraint Reconstruction

To mitigate missing modality induced representation shift and repair collapsed feature distributions, we design a HDCR. Unlike conventional pointwise matching objectives (e.g., MSE), HDCR is built on the principle of anchoring the statistical structure of the original modality in a reproducing kernel Hilbert space (RKHS), thereby forcing degraded representations back to their original distributional support.

3.2.1 Distributional Metric

HDCR adopts Maximum Mean Discrepancy (MMD) (Gretton et al., 2006) as its core optimization metric. Given the original feature distribution P and the reconstructed feature distribution Q, the squared MMD in an RKHS $\mathcal{H}$ is defined as:

$$D_{\mathcal{H}}(P,Q) = |\mathbb{E}x \sim P[\phi(x)] - \mathbb{E}y \sim Q[\phi(y)]|^2_{\mathcal{H}} \tag{1}$$

where $\phi(\cdot)$ is the feature map induced by a multi-bandwidth Gaussian kernel $k(x,y) = \sum_n \exp(-\frac{1}{\sigma_n}\|x-y\|^2)$. With multiple bandwidths, this metric captures distributional characteristics at different scales, enabling a finer manifold repair than MSE.

3.2.2 Bottom-up Hierarchical Constraints

HDCR utilizes a 2-layer Transformer backbone to generate reconstructed features $\hat{X}_m$ from the incomplete inputs F_m. To ensure that $\hat{X}_m$ preserves unimodal fidelity while maintaining cross-modal synergy, we propose a two-level objective:

Level-1: Unimodal Distribution Alignment. To retain modal specific semantic details, e.g., subtle facial micro-expressions or acoustic prosody cues. We independently align the original distribution P_m (from X_m) and the reconstructed distribution Q_m (from $\hat{X}_m$):

$$\mathcal{L}^m_{uni} = D_{\mathcal{H}}(P_m, Q_m) \quad m \in \{a,v,t\} \tag{2}$$

This term encourages the reconstruction to recover higher-order statistics within each modality, preventing the loss of discriminative information.

Level-2: Joint Distribution Consistency. Beyond unimodal integrity, we further impose a joint distribution constraint on the concatenated global representation to preserve inter-modal semantic relations and promote collaboration. Specifically, the reconstructed joint distribution Q_c is aligned with the original joint distribution P_c:

$$\mathcal{L}_{joint} = D_{\mathcal{H}}(P_c, Q_c) \tag{3}$$

where P_c and Q_c are formed from the original feature sequence $[X_t; X_v; X_a]$ and the reconstructed sequence $[\hat{X}_t; \hat{X}_v; \hat{X}_a]$, respectively.

The overall reconstruction loss of HDCR is a weighted sum of the two-level constraints, balanced by hyperparameters λ_1 and λ_2:

$$\mathcal{L}_{rec} = \sum_{m=1}^{M} \lambda_1 \mathcal{L}^m_{uni} + \lambda_2 \mathcal{L}_{joint} \tag{4}$$

With this hierarchical design, HDCR not only repairs local distribution collapse but also ensures that the global manifold structure of the joint representation converges toward the source domain before subsequent fusion, thereby reducing the missingness-induced domain gap at its root.The theoretical motivation for this hierarchical alignment, specifically regarding how it minimizes the generalization bound in missing-modality scenarios, is detailed in Appendix A.

3.3 Sample-level Uncertainty and Reliability Estimation

Although HDCR narrows the distribution gap from a global perspective, samples can be affected by missing perturbations to varying degrees, requiring the model to perceive the remaining instance-level shift. To this end, we propose the Sample-level Uncertainty and Reliability Estimation (SURE) mechanism. It captures distributional distortions in an emotion-anchored space and performs risk self-calibration via a Monotonic Weighting Network (MWN).

3.3.1 Emotional Subspace Projection

Under missing-modality conditions, the incomplete features $_m$ are easily corrupted by completion noise or scale fluctuations. Estimating reliability directly from F_m may trap the model in a feature magnitude pitfall, where the model mistakenly perceives

the amplitude drop caused by zero-filling as low difficulty rather than high uncertainty. To address this, we process the encoded features differently based on their semantic stability:

$$\begin{aligned} H_t &= \text{Enc}_t(F_t), \\ H_{\{v,a\}} &= \phi_{\{v,a\}}(F_{\{v,a\}}) \end{aligned} \tag{5}$$

where $\phi(\cdot)$consists of LayerNorm followed by a linear projection. As implemented in our framework, the textual branch $H_t \in \mathbb{R}^{T\times d_t}$ retains high-level semantic content, the visual and audio representations $H_v, H_a \in \mathbb{R}^{T\times d_{sent}}$ are mapped into a stable emotional subspace. This design aims to filter out emotion-irrelevant perturbations in a data-driven manner, providing a scale-decoupled and semantically focused basis for subsequent difficulty estimation.

3.3.2 Transfer Difficulty and Reliability Mapping

To quantify how far each sample deviates from the source-domain manifold, we propose the Transfer Difficulty Estimator (TDE). The TDE senses instance-level shifts by constructing a scale-insensitive descriptor from the projected features H_m. Specifically, we first aggregate and normalize the features to form the modality descriptor

$$z_m = Norm(Pooling(\tilde{H}m)) \tag{6}$$

where $Norm$ ensures that the descriptor's discriminative power stems from distributional shape rather than absolute energy. TDE then predicts a transfer-difficulty vector $\mathbf{d}$ by integrating cross-modal context:

$$\mathbf{d} = \text{Softplus}(MLP([z_t; z_v; z_a])) \in \mathbb{R}^3_{\geq 0} \tag{7}$$

where $\mathbf{d} = [d_t, d_v, d_a]^T$. This vector reflects the potential cost of moving each modality from its current missing state back to the complete distributional support, larger values indicate more severe shift.

To translate these abstract difficulty scores into intuitive fusion weights, we introduce the Monotonic Weighting Network (MWN). Following the heuristic that *higher difficulty implies lower reliability*, MWN produces reliability weights $\mathbf{w}$ via a temperature-calibrated mapping:

$$\mathbf{w} = \text{softmax}\left(-\frac{\mathbf{d}}{\tau}\right) \tag{8}$$

where $\mathbf{w} = [w_t, w_v, w_a]^T \in \mathbb{R}^3$, and the temperature τ acts as a risk calibrator. In extreme missing scenarios, dynamically adjusting the smoothness induced by τ helps prevent overconfident weight assignments, thereby mitigating the propagation of erroneous information during fusion.

3.4 Synergistic-Robust Dual-Stream Fusion

After obtaining the reliability weights $\mathbf{w}$, DEAR adopts a Synergistic-Robust Dual-Stream (SR-DS) fusion architecture to balance high-performance affective feature mining with decision robustness under extreme missingness. SR-DS consists of (i) an attention-based synergistic interaction stream and (ii) a linearly weighted robust fallback stream, with dynamic arbitration controlled by uncertainty-aware gating.

Synergistic Stream. The synergistic stream aims to compensate corrupted modalities using residual information from the available ones. Since text often carries dominant semantics in MSA, we use textual features as the main query. Let H_t denote the deep textual representation encoded by a Transformer, and H_v, H_a denote the visual and audio features. We explicitly modulate the contribution of non-text modalities in attention using reliability weights w_v, w_a, suppressing noise propagation from highly shifted signals:

$$\begin{aligned} Q &= H_t W_Q, \\ K &= [w_v \cdot (H_v W_K);\ w_a \cdot (H_a W_K)], \\ V &= [w_v \cdot (H_v W_V);\ w_a \cdot (H_a W_V)], \\ H_{syn} &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \end{aligned} \tag{9}$$

where W_Q, W_K, W_V are learnable projection matrices. By scaling the K and V of each modality by its corresponding reliability weight, the model dynamically suppresses the influence of unreliable modalities at the feature-interaction level.

The final synergistic representation H_{syn} is passed through a regression head to produce the synergistic prediction y_{syn}.

Robust Stream. When multiple modalities collapse severely,complex non-linear interactions in the SS may amplify uncertainty and lead to catastrophic failure. The robust stream therefore adopts a conservative strategy, avoiding noise amplification by directly aggregating modal specific predictors with reliability weights:

$$y_{rob} = w_t \cdot \hat{y}_t + w_v \cdot \hat{y}_v + w_a \cdot \hat{y}_a \tag{10}$$

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
MISA	29.85	33.08	71.49/70.33	71.28/70.00	1.085	0.524	40.84	39.39	71.27/75.82	63.85/68.73	0.780	0.503
Self-MM	29.55	34.67	70.51/69.26	66.60/67.54	1.070	0.512	44.70	45.38	73.89/77.42	68.92/72.31	0.695	0.498
MMIM	31.30	33.77	69.14/67.06	66.65/64.04	1.077	0.507	40.75	41.74	73.32/75.89	68.72/70.32	0.739	0.489
CENET	30.38	37.25	71.46/67.73	68.41/64.85	1.080	0.504	**47.18**	47.83	74.67/77.34	70.68/74.08	0.685	0.535
TETFN	30.30	34.34	69.76/67.68	65.69/63.29	1.087	0.507	30.30	47.70	69.76/67.68	65.69/63.29	1.087	0.508
TFR-Net	29.54	34.67	68.15/66.35	61.73/60.06	1.200	0.459	46.83	34.67	73.62/77.23	68.80/71.99	0.697	0.489
ALMT*	31.99	34.97	71.94/70.79	72.04/70.80	1.067	0.505	44.82	45.88	78.44/77.52	77.56/77.16	0.669	0.586
LNLN*	31.57	34.66	72.58/71.46	72.44/71.24	1.073	0.507	43.83	44.49	78.16/76.95	76.65/76.94	0.666	0.589
P-RMF*	28.36	35.96	71.10/70.39	71.13/70.52	1.117	0.487	44.41	44.79	79.16/79.06	78.51/77.97	0.668	0.581
DAR	**34.47**	**38.65**	**73.18**/71.60	73.15/71.51	1.069	0.520	47.01	**48.02**	78.14/77.48	77.51/77.44	0.665	0.583
DEAR	33.80	36.40	72.43/**72.08**	**73.70/72.76**	**1.049**	**0.528**	45.46	46.65	**79.26/79.35**	**79.03/78.35**	**0.657**	**0.591**

Table 1: Robustness comparison of the overall performance on MOSI and MOSEI datasets. Note: * represents the result coming from our re-run. Rest of the results are copy from LNLN (Zhang et al., 2024) and DAR (Yang et al., 2025). Bold and underline mean the best and second-best results, respectively.

where $\hat{y}_m$ is the prediction produced by an independent unimodal regressor. By avoiding complex cross-modal propagation, this stream is highly resilient and interpretable under extreme missingness, providing a lower-bound support for prediction quality.

DEAR finally performs instance-wise arbitration between the two streams via a learnable gate g, which is driven by the weight entropy $H(\mathbf{w})$ and the overall difficulty. The weight entropy reflects the model's overall confidence: a low entropy implies clear reliability, while a high entropy indicates severe shift across all modalities. The final prediction is formulated as:

$$\begin{aligned} g &= \sigma(\text{MLP}([H_t^{pool}; \mathbf{d}; H(\mathbf{w})])), \\ \hat{y} &= g \cdot y_{syn} + (1-g) \cdot y_{rob} \end{aligned} \tag{11}$$

3.5 Self-supervised Calibration and Task Objectives

To ensure that the difficulty vector $\mathbf{d} \in \mathbb{R}^3_{\geq 0}$ predicted by SURE is physically meaningful, we introduce a self-supervised calibration task. We define the reconstruction deviation $e_m = \|X_m - \hat{X}_m\|_2^2$ produced by HDCR as a proxy for the ground-truth representation shift. TDE is then encouraged to approximate this deviation:

$$\mathcal{L}_{cal} = \frac{1}{M} \sum_{m \in \{t,v,a\}} (e_m - d_m)^2 \tag{12}$$

Through this regression, SURE learns to sense the magnitude of information loss directly from incomplete inputs F_m without accessing the original data during inference.

The final training objective is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \alpha \mathcal{L}_{aux} + \mathcal{L}_{rec} + \beta \mathcal{L}_{cal} \tag{13}$$

where α, β are hyperparameters.

Method	Acc-5	Acc-3	Acc-2	F1	MAE↓	Corr
MISA	31.53	56.87	72.71	66.30	0.539	0.348
Self-MM	32.28	56.75	72.81	68.43	0.508	0.376
MMIM	31.81	52.76	69.86	66.21	0.544	0.339
CENET	22.29	53.17	68.13	57.90	0.589	0.107
TETFN	33.42	56.91	73.58	68.67	0.505	0.387
TFR-Net	26.52	52.89	68.13	58.70	0.661	0.169
ALMT*	30.80	52.81	71.28	69.25	0.532	0.336
LNLN*	30.15	55.34	73.69	70.26	0.511	0.391
P-RMF*	29.69	56.35	71.19	66.78	0.539	0.343
DEAR	**34.64**	**57.57**	**74.64**	**70.79**	**0.503**	**0.423**

Table 2: Robustness comparison of the overall performance on SIMS dataset. Note: * represents the result coming from our re-run. Rest of the results are copy from LNLN (Zhang et al., 2024). Bold and underline mean the best and second-best results, respectively.

4 Experiments and Analysis

In this section, we provide a comprehensive analysis of the results between DEAR and previous SOTA methods on MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018) and SIMS (Yu et al., 2020) datasets. Further experimental details and baseline descriptions are provided in Appendix B.

4.1 Main Result Comparison

Tables 1 and 2 report the evaluation results of DEAR against a range of recent SOTA baselines on three datasets. More detailed test results are provided in Appendix C.9.

On MOSI, DEAR achieves the best F1, which we attribute to the SR-DS dual-stream fusion that

Method	MOSI						SIMS					
	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr	Acc-5	Acc-3	Acc-2	F1	MAE↓	Corr
DEAR	**33.80**	**36.40**	**72.43/72.08**	**73.70/72.76**	**1.049**	**0.528**	**36.64**	**57.57**	**74.64**	**70.79**	**0.503**	**0.423**
w/o $\mathcal{L}_{uni}$	32.31	35.36	71.39/71.45	72.29/71.40	1.055	0.512	34.54	56.34	73.23	69.43	0.516	0.412
w/o $\mathcal{L}_{joint}$	31.89	35.04	71.13/70.87	72.32/71.23	1.061	0.508	35.67	56.11	72.89	69.29	0.522	0.409
w/o HDCR	31.13	34.54	70.78/70.66	71.87/71.44	1.134	0.503	34.43	55.53	72.56	69.01	0.530	0.404
w/o SURE	33.21	35.21	71.36/70.98	72.55/71.67	1.174	0.515	36.04	56.23	73.12	69.89	0.523	0.412
w/o SS	31.37	34.68	70.69/71.23	71.84/70.36	1.066	0.512	34.89	55.21	72.22	68.93	0.521	0.403
w/o RS	32.54	34.96	71.23/71.78	72.46/71.08	1.058	0.518	35.54	56.33	73.74	69.86	0.517	0.411

Table 3: Ablation results on MOSI and SIMS datasets. w/o HDCR means using MSE as the reconstruction optimizer, SS denotes the synergistic Stream, and RS denotes the Robust Stream. Ablation experiments for MOSEI are presented in Appendix C.4.

adapts prediction behavior to input quality, yielding more stable decisions than fixed aggregation baselines. While it is slightly behind DAR on Acc-7/Acc-5, DEAR attains 72.76% on F1 and the best Corr, indicating that HDCR effectively aligns distributions and improves the numerical fidelity of reconstructed features, thereby reducing prediction collapse. On MOSEI, the large scale and severe class imbalance limit all methods on Acc-7/Acc-5, yet DEAR remains leading on overall metrics such as Acc-2, F1, and MAE, further validating HDCR's manifold repair under high missing rates.

In Table 2, DEAR performs best in all metrics, benefiting from the synergy between SURE and the dual stream predictor, where SURE captures modality transfer difficulty and sharpens sentiment boundaries in multi-class settings. Overall, DEAR is robust for binary tasks and consistently strong for multi-class and regression, highlighting the generality of its distribution-aware design under missing-modality uncertainty.

4.2 Ablation Study

To investigate the contribution of the core modules of DEAR, we conducted ablation experiments on MOSI and SIMS, our analysis focuses on both the reconstruction optimization strategy and the reliability-driven prediction scheme.

As shown in Table 3, performance drops without $\mathcal{L}_{uni}$ or $\mathcal{L}_{joint}$, proving that capturing both individual and collaborative distributions is vital for alignment. Replacing HDCR with MSE loss causes the sharpest decline, validating that explicit distributional constraints outperform simple pixel-level completion in mitigating semantic shifts.Prediction. Ablating SURE reduces stability as the model loses the ability to evaluate reconstruction fidelity and rebalance dual-stream weights. Further, the sub-optimal results without SS or RS confirm that dynamic arbitration between synergy and robustness is key to maintaining consistent reliability across varying missing intensities.

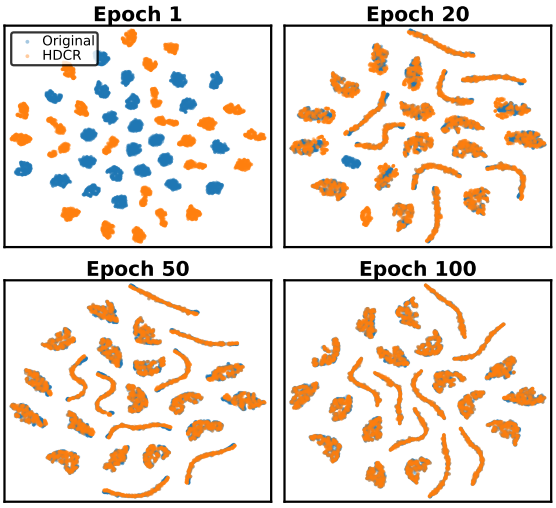


Figure 3: T-SNE (Hinton and Roweis, 2002) visualization of the distribution of the raw features and the joint features obtained by the HDCR module.

4.3 The Potency of HDCR Module

Fig. 3 illustrates the distribution evolution of raw features versus HDCR-enhanced features across different training stages using t-SNE mapping. At the early stage of training, the two feature sets exhibit a pronounced distribution shift: the orange points are largely separated from the blue ones, indicating the semantic deviation caused by missing modalities. As training proceeds, under HDCR's distribution-consistency constraints, the orange clusters are progressively pulled toward the original semantic centers. By the late stage, the two distributions become highly topologically aligned and largely overlap. These results provide strong evidence that HDCR can explicitly correct the fea-

Dataset	Method	$\{L\}$	$\{A\}$	$\{V\}$	$\{L, A\}$	$\{L, V\}$	$\{V, A\}$	Avg.
MOSI	GCNet‡ (Lian et al., 2023)	80.91	65.07	58.70	84.73	83.58	70.02	73.84
	DicMoR (Wang et al., 2023c)	84.40	60.80	60.20	85.50	85.40	63.50	73.30
	IMDer (Wang et al., 2023d)	84.70	62.20	60.80	85.30	85.40	63.40	73.63
	MPLMM (Guo et al., 2024)	80.31	63.65	63.74	81.09	81.19	65.41	72.57
	CorrKD‡ (Li et al., 2024c)	81.20	66.52	60.72	83.56	82.41	73.74	74.69
	HRLF‡ (Li et al., 2024b)	83.36	69.47	64.59	83.82	83.56	**75.62**	76.74
	DEAR	**85.85**	**71.67**	**68.57**	**85.83**	**84.89**	74.32	**78.52**
MOSEI	GCNet‡ (Lian et al., 2023)	80.52	66.54	61.83	81.96	81.15	69.21	73.54
	DicMoR (Wang et al., 2023c)	84.30	63.60	60.40	84.90	84.90	64.40	73.75
	IMDer (Wang et al., 2023d)	84.50	60.60	63.60	85.10	**85.00**	63.50	73.72
	MPLMM (Guo et al., 2024)	79.17	68.71	69.40	80.43	80.13	69.91	74.62
	CorrKD‡ (Li et al., 2024c)	80.76	66.09	62.30	81.74	81.28	71.92	74.02
	HRLF‡ (Li et al., 2024b)	82.05	69.32	64.90	82.62	81.09	73.80	75.63
	DEAR	**84.73**	**70.38**	**69.71**	**85.43**	84.72	**73.97**	**78.16**

Table 4: Quantitative F1 results under six possible missing modality cases. For example, "$\{L\}$" means language modality is available while visual and audio are missing. ‡ denotes results copied from (Li et al., 2024b), other results are reporting from the original paper.

ture gap and improve semantic recovery for missing modalities, yielding more discriminative and robust representations for subsequent tasks.

4.4 Robustness to Inter-modality Missingness

To evaluate generalizability, we simulate six cross-modal missing scenarios in Table 4. Our findings are summarized as follows: (1) Across all methods, the language modality ($\{L\}$) consistently remains the most informative, providing a higher performance baseline than audio or visual modalities. (2) Our framework achieves state-of-the-art F1 scores in most missing patterns, with an average improvement on MOSEI over the HRLF baseline. This demonstrates that DEAR's reliability driven arbitration effectively mitigates information loss and safeguards prediction stability.

4.5 Analysis of Decision Reliability

We evaluate DEAR's decision mechanism across 10 missingness levels averaged over three seeds. As shown in Fig. 4, the steady rise in weight confirms that the SURE module effectively captures uncertainty from distribution collapse. Consequently, the gating value g exhibits a distinct decline, validating that SR-DS proactively shifts from cross-modal synergy to a robust aggregation strategy as information becomes scarce. The weight heatmap further reveals that the system adaptively suppresses corrupted audio-visual weights to 0.15 while re-centering emphasis on the stable textual modality. This coupled evolution demonstrates DEAR's ability to filter noise and maintain a reliable performance lower bound under extreme conditions.

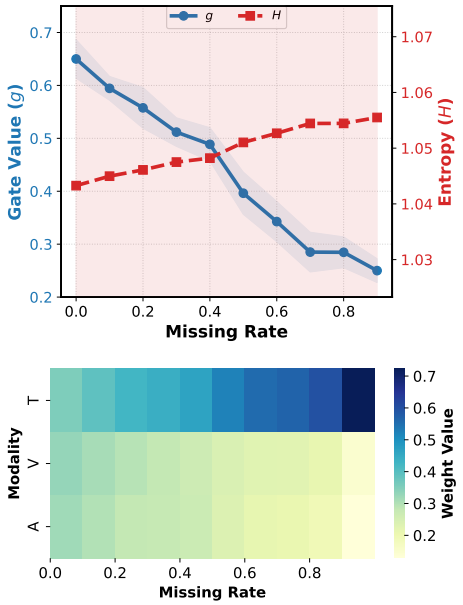


Figure 4: Analysis of decision reliability. (a) Dynamics of gate value and weight entropy across varying missing rates. (b) Visualization of modality reliability weights under different missing levels.

5 Conclusion

In this work, we propose DEAR, an uncertainty-aware framework for robust multimodal learning. HDCR imposes distribution-consistency constraints during reconstruction to alleviate missing-modality–induced semantic shift, while SURE estimates risk to guide SR-DS in adaptively switching between synergistic fusion and robust aggregation. Experiments and visualizations demonstrate that DEAR delivers leading performance on multiple benchmarks and maintains reliable predictions by suppressing noise under extreme conditions.

Limitation

While the proposed DEAR framework demonstrates significant robustness in multimodal sentiment analysis, several limitations should be noted. Firstly, due to constraints in computational resources and time, we referred to the officially reported results of certain baselines from recent state-of-the-art studies; to ensure a fair comparison, we strictly adhered to identical datasets and experimental protocols. Secondly, although the HDCR module effectively mitigates semantic shifts, its computational overhead is slightly higher than simple MSE-based completion when processing complex feature manifolds, suggesting a need for more efficient alignment algorithms. Finally, while we evaluated DEAR across various missingness intensities, real-world environments often involve more intricate joint missingness patterns. Furthermore, due to the stochastic nature of missing data noise, model performance can be subject to inherent uncertainty across certain metrics. This highlights the importance of balancing stability and performance under volatile noise levels to enhance the practical reliability of MSA systems.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79:151–175.

Bingfeng Chen, Qihan Ouyang, Yongqi Luo, Boyan Xu, Ruichu Cai, and Zhifeng Hao. 2024. S^2gsl: Incorporating segment to syntactic enhanced graph structure learning for aspect-based sentiment analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13366–13379.

Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. 2024. Lggnet: Learning from local-global-graph representations for brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 35:9773–9786.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. 2006. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3394–3402.

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*, pages 1726–1736.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Ye He, Shihao Zou, Yuzhe Chen, and Xianying Huang. 2024. C^3lpgcn:integrating contrastive learning and cooperative learning with prompt into graph convolutional network for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL*, pages 3237–3247.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. In *Advances in neural information processing systems*.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, pages 10944–10956.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. 2024a. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10074–10082.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. 2024b. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *Advances in Neural Information Processing Systems*.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024c. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6631–6640.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:8419–8432.

Jingjun Liang, Ruichen Li, and Qin Jin. 2020. Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2852–2861.

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

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2554–2562.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 6892–6899.

Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Mohammed E. Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8821–8831.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30st ACM International Conference on Multimedia*, pages 3722–3729.

Wenjin Tian, Xianying Huang, and Shihao Zou. 2025. Multi-condition guided diffusion network for multimodal emotion recognition in conversation. In *Findings of the Association for Computational Linguistics: NAACL*, pages 3215–3227.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

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

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2023b. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia,*, 25:4909–4921.

Yuanzhi Wang, Zhen Cui, and Yong Li. 2023c. Distribution-consistent modal recovering for incomplete multimodal learning. In *IEEE International Conference on Computer Vision*, pages 21968–21977.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023d. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, pages 17117–17128.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651.

Mingzheng Yang, Kai Zhang, Yuyang Ye, Yanghai Zhang, Runlong Yu, and Min Hou. 2025. Decoupling and reconstructing: A multimodal sentiment analysis framework towards robustness. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 6803–6811.

Shuo Yin and Guoqiang Zhong. 2024. Textgt: A double-view graph transformer on text for aspect-based sentiment analysis. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, pages 19404–19412.

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

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10790–10797.

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

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2024. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246.

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

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

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767.

Xiangmin Zhang, Wei Wei, and Shihao Zou. 2025. Modal feature optimization network with prompt for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING*, pages 4611–4621.

Sicheng Zhao, Zhenhua Yang, Henglin Shi, Xiaocheng Feng, Lingpengkun Meng, Bing Qin, Chenggang Yan, Jianhua Tao, and Guiguang Ding. 2025. Sdrs: Sentiment-aware disentangled representation shifting for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–13.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 22123–22138.

A Theoretical Insights on Generalization Bound

In this section, we provide a rigorous theoretical characterization of DEAR's generalization performance under the domain adaptation framework in statistical learning theory, and derive how HDCR reduces the target-domain risk by shrinking the divergence term.

A.1 Preliminaries and Error Decomposition

Let $\mathcal{X}$ denote the input feature space and $\mathcal{H}$ the hypothesis space. We define the expected risks on the source domain (complete modalities) and the target domain (missing modalities) as $R_{\mathcal{S}}(h)$ and $R_{\mathcal{T}}(h)$, respectively. Following the generalization bound of Ben-David (Ben-David et al., 2010), the target risk can be decomposed as:

$$R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda \quad (14)$$

where

$$\lambda = \min_{h \in \mathcal{H}} \left[R_{\mathcal{S}}(h) + R_{\mathcal{T}}(h) \right]$$

denotes the optimal joint error under an ideal hypothesis. In missing-modality settings, distribution collapse caused by signal absence makes the second term: the $\mathcal{H}$-divergence $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$, the primary challenge. It is defined as:

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2 \sup_{h \in \mathcal{H}} [&\Pr_{x \sim \mathcal{S}}[h(x) \neq 0] \\ &- \Pr_{x \sim \mathcal{T}}[h(x) \neq 0]] \end{aligned} \quad (15)$$

A.2 Hierarchical Shrinkage of the Risk Bound via HDCR

Conventional MSE optimization only enforces point-wise matching of second-order moments, i.e., $\mathbb{E}[x_{\mathcal{S}}] \approx \mathbb{E}[x_{\mathcal{T}}]$, but this does not guarantee distributional overlap in RKHS. By contrast, DEAR introduces HDCR with hierarchical distribution constraints, which directly suppress the $d_{\mathcal{H}\Delta\mathcal{H}}$ term.

(1) Recovering unimodal distribution manifolds. HDCR first aligns unimodal marginal distributions by minimizing

$$\mathcal{L}_{uni}^{m} = D_{\mathcal{H}}(P_m, Q_m) \quad (16)$$

Theoretically, MMD provides a strict upper bound on the $\mathcal{H}$-divergence. With the kernel mapping $\phi(\cdot)$, features are projected into a high-dimensional space such that:

$$d_{\mathcal{H}\Delta\mathcal{H}}(P_m, Q_m) \leq 2 \cdot \text{MMD}_{\mathcal{H}}(P_m, Q_m) + C \quad (17)$$

This implies that optimizing Eq.(16) effectively tightens the unimodal generalization gap and repairs the distributional support of low-level features.

(2) Closing the joint distribution gap. To further address the distribution shift under cross-modal interaction, $\mathcal{L}_{joint}$ constrains the joint distributions P_c and Q_c. Let $h \circ f$ denote the composite classifier; its complexity in the joint space may cause the λ term to vary. HDCR mitigates this by optimizing:

$$\min_{\theta_{HDCR}} \sum_{m=1}^{M} \lambda_1 D_{\mathcal{H}}(P_m, Q_m) + \lambda_2 D_{\mathcal{H}}(P_c, Q_c) \quad (18)$$

which anchors distributions from bottom-up low-level features to top-down joint semantics. Consequently, under missing patterns, the RKHS distance between reconstructed features Q_c and original features P_c satisfies $\text{dist}(P_c, Q_c) \to 0$.

B Experiments Details and Baselines

B.1 Datasets and Evaluation Metrics

Datasets. The MOSI dataset contains 2,199 multimodal samples integrating visual, audio, and textual modalities. Each sample is annotated with a sentiment score ranging from -3 (strongly negative) to +3 (strongly positive). The MOSEI dataset is a larger-scale multimodal sentiment dataset with the same modalities and sentiment scoring scheme as MOSI, offering more diverse and extensive samples. The SIMS dataset is a Chinese multimodal sentiment dataset comprising 2,281 video clips collected from various movies and television dramas. Each sample is manually annotated with a sentiment score ranging from -1 (negative) to +1 (positive). The statistics of these datasets are summarized in Table 5.

Evaluation metrics. We report the binary classification accuracy (Acc-2), the F1 score associated with Acc-2, three-class accuracy (Acc-3), the seven-class accuracy (Acc-7), mean absolute error (MAE) and the correlation of the model's prediction with humans (Corr). For Acc-2, we calculated accuracy and F1 in two ways: negative/non-negative and negative/positive on the MOSI and

Dataset	Speaker	Video clip	Train	Valid	Test	Language
MOSI	93	2199	1284	229	686	English
MOSEI	1000	22856	16326	1871	4659	English
SIMS	474	2281	1368	456	457	Chinese

Table 5: Dataset statistics.

MOSEI datasets, respectively. For the SIMS dataset, we report Acc-2, Acc-3, the five-class accuracy (Acc-5), F1 Score, MAE and Corr.

Experimental details follow the LNLN (Zhang et al., 2024) experimental setup, with both training and testing conducted on a single GeForce RTX 4090 GPU.

B.2 Baselines

- **MISA** (Hazarika et al., 2020) projects each modality into two distinct subspaces. The first is modality-invariant, where shared features across modalities are learned to reduce modality gaps. The second is modality-specific, capturing private characteristics unique to each modality.
- **Self-MM** (Yu et al., 2021) utilizes a self-supervised label generation module to obtain unimodal supervision.
- **MMIM** (Han et al., 2021) maximizes mutual information between unimodal inputs (intermodal) and between multimodal fused representations and unimodal inputs.
- **TFR-Net** (Yuan et al., 2021) employs intra- and inter-modal attention-based extractors to learn robust representations from modality sequences.
- **CENET** (Wang et al., 2023b) enhances text representations by integrating visual and acoustic information into a language model.
- **TETFN** (Wang et al., 2023a) learns text-guided pairwise cross-modal mappings to obtain effective unified multimodal representations.
- **ALMT** (Zhang et al., 2023) introduces an Adaptive Hyper-modality Learning module to suppress irrelevant/conflicting features from visual and acoustic inputs using multi-scale language guidance.
- **LNLN** (Zhang et al., 2024) ensures the quality of dominant modality representations through specialized modules, enhancing robustness under various noise conditions.
- **P-RMF** (Zhu et al., 2025) learns a robust proxy modality by modeling modality uncertainty in a Gaussian latent space and enhancing it via dynamic cross-modal injection.
- **DAR** (Yang et al., 2025) decouples features into common/independent components and reconstructs them separately to handle missing or unaligned modalities.
- **DicMoR** (Wang et al., 2023c) proposes a category-specific flow-based modality recovery method that transforms cross-modal distributions conditioned on sample class.
- **IDMer** (Wang et al., 2023d) exploits the score-based diffusion model that maps the input Gaussian noise into the desired distribution space of the missing modalities and recovers missing data abided by their original distributions.
- **MPLMM** (Guo et al., 2024) introduces three types of prompts generation, missing signal, and missing type which help generate missing modality features and enhance intra- and inter-modal learning.
- **CorrKD** (Li et al., 2024c) reconstructs missing semantics by transferring comprehensive knowledge containing cross-sample correlations.
- **HRLF** (Li et al., 2024b) introduces a fine-grained representation factorization module that decomposes modalities into emotion-relevant representations via cross-modal translation and semantic reconstruction.

C Additional Experiments and Analysis

C.1 Effect of Hyperparameters and Layers

We conducted hyperparameter sensitivity experiments on MOSI and SIMS to evaluate the robustness of DEAR across varying configurations of hierarchical reconstruction parameters (λ_1, λ_2) and loss weights (α, β). As shown in Table 6, the model achieves optimal performance with $\lambda_1 = 0.3, \lambda_2 = 0.7, \alpha = 0.8$, and $\beta = 0.3$. Notably, DEAR maintains consistently high performance with minimal fluctuations under different parame-

MOSI						
λ_1, λ_2, α, β	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
0.3, 0.7, 0.8, 0.3	33.80	**36.40**	**72.43 / 72.08**	**73.70 / 72.76**	**1.049**	**0.528**
0.2, 0.8, 0.7, 0.2	**33.94**	36.25	72.41 / 71.84	72.79 / 71.87	1.055	0.523
0.4, 0.6, 0.6, 0.1	32.76	36.33	72.20 / 71.69	72.46 / 71.72	1.078	0.511
0.5, 0.5, 0.5, 0.4	32.35	35.23	72.20 / 70.69	72.43 / 71.22	1.065	0.514
SIMS						
λ_1, λ_2, α, β	Acc-5	Acc-3	Acc-2	F1	MAE↓	Corr
0.3, 0.7, 0.8, 0.3	**34.64**	**57.57**	**74.64**	**70.79**	**0.503**	**0.423**
0.2, 0.8, 0.7, 0.2	33.75	56.78	73.20	70.04	0.489	0.417
0.4, 0.6, 0.6, 0.1	33.54	55.86	73.23	69.86	0.485	0.414
0.5, 0.5, 0.5, 0.4	32.61	56.11	73.72	70.20	0.477	0.389

Table 6: Hyperparameter sensitivity analysis.

MOSI						
Layer	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
2	**33.80**	**36.40**	**72.43 / 72.08**	**73.70 / 72.76**	**1.049**	0.528
4	32.48	35.66	70.78 / 71.59	71.90 / 71.60	1.067	**0.548**
6	31.69	35.35	70.23 / 70.98	71.03 / 71.24	1.089	0.510
SIMS						
Layer	Acc-5	Acc-3	Acc-2	F1	MAE↓	Corr
2	**34.64**	**57.57**	**74.64**	**70.79**	**0.503**	**0.423**
4	33.23	56.71	73.48	69.65	0.511	0.417
6	32.83	55.58	73.57	69.26	0.515	0.404

Table 7: Effect of reconstruction layers on model performance.

ter combinations, demonstrating its superior stability and robustness in diverse multimodal sentiment analysis scenarios.

We investigated the impact of the number of layers within the HDCR module on model performance. As shown in Table 7, the framework achieves its peak performance across most metrics on both MOSI and SIMS when the number of layers is set to 2. Increasing the depth to 4 or 6 layers leads to a marginal decline in accuracy and F1-scores. This suggests that a moderate depth is sufficient for capturing the distributional features of missing modalities, whereas excessive layers may introduce parameter redundancy or hinder the optimization of distributional consistency. Consequently, we adopt a 2-layer Transformer configuration to balance high-fidelity reconstruction with computational efficiency.

C.2 Efficiency Analysis

Model	Parameters	Time / Epoch
P-RMF	117 M	16 s
LNLN	116 M	7.8 s
DEAR	114 M	7.6 s

Table 8: Results of Computational Efficiency.

Table 8 shows the computational overhead of the proposed DEAR model on the MOSI dataset, and compares it with the SOTA model P-RMF and LNLN for incomplete modality under the same experimental setup. Compared to LNLN and P-RMF, our method achieves a shorter runtime per epoch and maintaining a smaller number of parameters, indicating that its model structure is more optimized and computationally efficient.

	Acc-7	Acc-5	MAE	Corr	Time/Epoch
MMD	**33.80**	**36.40**	**1.049**	**0.528**	**7.6 s**
WD	32.84	35.64	1.148	0.492	33 s

Table 9: Comparison between MMD and Wasserstein distance (WD) for distribution alignment on MOSI.

C.3 MMD vs. Wasserstein Distance for Alignment

To further justify this choice, we replace MMD with a differentiable Wasserstein approximation (Sinkhorn) under the same experimental setting and compare both task performance and training efficiency on MOSI. As shown in Table 9, MMD consistently achieves better results than WD, improving Acc-7 from 32.84 to 33.80 and Acc-5 from 35.64 to 36.40. At the same time, MMD is substantially more efficient, reducing the per-epoch training time from 33 s to 7.6 s. These results indicate that MMD provides a more favorable effectiveness-efficiency trade-off for distribution alignment in our framework, while WD remains useful as an intuitive metric for visualization.

C.4 Ablation Study on MOSEI

To further verify the generalizability of DEAR, we provide ablation results on the larger-scale MOSEI dataset in Table 10. The results show that removing either $\mathcal{L}_{uni}$ or $\mathcal{L}_{joint}$ leads to performance degradation, particularly in Acc-2, which underscores the necessity of capturing both individual and collaborative distributions for high-fidelity alignment. Notably, replacing HDCR with MSE loss causes a sharp decline in the correlation coefficient (Corr) from 0.591 to 0.500, validating that explicit distributional constraints are superior in mitigating semantic shifts within high-dimensional feature manifolds. Regarding the prediction module, the removal of SURE or either stream (SS/RS) results in sub-optimal performance. Specifically, the increased MAE and decreased F1-score upon ablating the RS path further confirm that SR-DS's dynamic arbitration is crucial for safeguarding the decision lower bound and maintaining consistent reliability under uncertainty.

C.5 Reliability under Feature-space Quality Degradation

To examine whether the reliability scores predicted by SURE reflect the actual usefulness of each modality, we perform a controlled feature-space degradation experiment. Since MOSI/MOSEI provide pre-extracted features rather than raw signals, we simulate modality corruption by injecting Gaussian noise with increasing variance into the visual modality during evaluation, while keeping the other modalities unchanged. As shown in Table 11, the reliability weight assigned to the visual modality (w_V) decreases strictly and monotonically as the noise level increases, from 0.393 at $\sigma = 0.0$ to 0.352 at $\sigma = 2.0$, yielding a Spearman correlation of $\rho = -1.0$. Meanwhile, the overall task performance also degrades consistently (Acc-7 drops from 43.81 to 39.87). This result indicates that SURE does not act as a static or heuristic weighting scheme; instead, it responds adaptively to modality degradation and captures the effective information content and uncertainty of the corrupted modality.

C.6 Ablations for Module Combinations

To further examine whether the gains of DEAR arise from genuine collaboration among modules rather than isolated component effects, we conduct additional multi-module combination ablations on MOSI. Specifically, removing HDCR replaces the distribution-level alignment with a standard MSE reconstruction loss, removing SURE uses uniform modality weights instead of dynamic reliability estimation, and removing DS degenerates the dual-stream architecture into a single-stream variant. As shown in Table 12, jointly removing any two components consistently leads to substantial degradation across all metrics. For example, removing HDCR+SURE decreases Acc-2 from 72.43/72.08 to 70.22/70.78, lowers F1 from 73.70/72.76 to 71.68/71.10, and increases MAE from 1.049 to 1.168. Similar performance drops are observed for w/o (HDCR+DS) and w/o (SURE+DS). These results indicate that the three components are not merely additive; rather, they form a complementary pipeline in which HDCR reduces reconstruction bias, SURE estimates modality reliability, and DS further suppresses noisy information through adaptive arbitration. Disabling any two of them disrupts this collaboration and results in markedly inferior performance.

	MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
DEAR	**45.46**	**46.65**	**79.26/79.35**	**79.03/78.35**	**0.657**	**0.591**
w/o $\mathcal{L}_{uni}$	45.21	46.58	78.33/78.45	78.51/78.03	0.671	0.579
w/o $\mathcal{L}_{joint}$	45.05	46.89	78.10/78.22	78.14/77.77	0.674	0.575
w/o HDCR	44.73	46.11	78.03/78.19	78.16/77.54	0.681	0.500
w/o SURE	45.11	46.76	78.12/78.24	78.69/78.15	0.663	0.587
w/o SS	43.82	45.24	77.68/77.79	77.87/77.35	0.689	0.567
w/o RS	44.23	46.47	78.13/78.25	78.38/78.10	0.676	0.578

Table 10: Ablation results of intra-modality missingness case on MOSEI datasets. w/o HDCR means using MSE as the reconstruction optimizer, SS denotes the synergistic Stream, and RS denotes the Robust Stream.

σ	w_L	w_V	w_A	Acc-7
0.0	0.217	0.393	0.390	43.81
0.5	0.221	0.381	0.387	42.07
1.0	0.239	0.363	0.376	40.92
2.0	0.262	0.352	0.366	39.87

Table 11: Reliability analysis under feature-space degradation on MOSI. Gaussian noise with increasing variance is injected into the visual modality during evaluation.

C.7 Modality Ablation Analysis

To quantify the contribution of each modality in DEAR, we conduct modality ablations on MOSI by removing one modality at a time. The results are reported in Table 13. As expected, removing the language modality causes the most significant degradation across all metrics, with Acc-2 dropping from 72.43/72.08 to 58.68/58.90 and Corr decreasing from 0.528 to 0.121. This is consistent with a common property of MSA, where text usually provides the dominant semantic signal. Nevertheless, removing the visual or audio modality also leads to consistent performance drops, e.g., MAE increases from 1.049 to 1.101 and 1.089, respectively. These results indicate that DEAR effectively exploits all three modalities, while appropriately relying more on text as the most informative source. In other words, the framework is not merely driven by language alone; instead, it benefits from complementary multimodal cues whenever they are available.

C.8 Query Choice and Fusion Comparison

We compare different query choices and representative fusion strategies on MOSI. Text is used as the query in DEAR because it usually provides the most stable semantic anchor under missing-modality settings. Table 14 shows that using text as query outperforms both audio-as-query and visual-as-query variants, and also exceeds standard fusion strategies such as Concat+SelfAttn and Gated Fusion across all metrics. This suggests that the benefit does not arise from heuristic preference, but from the interaction between a semantically reliable query and the proposed reliability-aware fusion mechanism.

C.9 Robustness to Intra-modality Missingness Details

To evaluate the stability of DEAR across different levels of data sparsity, we conducted stress tests under missing rates r ranging from 0.1 to 0.9. Experimental results are shown in Tables 15, 16, and 17. While performance metrics naturally decline as r increases, DEAR exhibits remarkable resilience across all three datasets. Even in extreme scenarios (e.g., $r = 0.7$), the Acc-2 on SIMS remains high at 71.05%, demonstrating the model's ability to extract critical affective features from information-sparse inputs. In terms of average metrics, DEAR achieves an excellent balance, particularly with low MAE and stable Corr values. This validates that our HDCR and SR-DS mechanisms effectively counteract the uncertainty induced by random missingness, ensuring practical reliability in volatile real-world environments.

	MOSI					
	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
DEAR	**33.80**	**36.40**	**72.43/72.08**	**73.70/72.76**	**1.049**	**0.528**
w/o HDCR+SURE	30.78	34.24	70.22/70.78	71.68/71.10	1.168	0.513
w/o HDCR+DS	30.11	33.90	70.05/70.92	71.34/71.26	1.177	0.505
w/o SURE+DS	31.21	34.73	71.11/71.23	71.41/71.00	1.182	0.499

Table 12: Multi-module combination ablations on MOSI dataset.

	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
DEAR	**33.80**	**36.40**	**72.43/72.08**	**73.70/72.76**	**1.049**	**0.528**
w/o language	20.54	19.89	58.68/58.90	60.78/62.89	1.545	0.121
w/o visual	32.11	34.34	71.78/71.23	72.44/71.19	1.101	0.510
w/o audio	32.65	35.31	71.98/71.77	72.76/71.76	1.089	0.521

Table 13: Modality ablation study on MOSI.

	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
DEAR	**33.80**	**36.40**	**72.43/72.08**	**73.70/72.76**	**1.049**	**0.528**
Audio as Query	30.45	34.89	70.34/69.98	71.13/71.19	1.353	0.477
Visual as Query	31.78	35.77	71.23/70.48	71.56/70.78	1.332	0.487
Concat+SelfAttn	28.43	33.71	68.89/69.20	69.77/69.80	1.534	0.332
Gated Fusion	31.23	34.45	71.09/69.79	70.22/70.10	1.228	0.458

Table 14: Comparison of different query choices and fusion strategies on MOSI.

Missing Rate r	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
0.0	43.81	46.20	85.59/82.41	84.92/83.80	0.751	0.778
0.1	42.19	45.35	80.58/80.47	82.52/81.42	0.842	0.724
0.2	38.27	44.44	78.62/78.43	80.37/80.18	0.860	0.666
0.3	37.54	40.77	77.47/77.41	78.35/77.35	0.924	0.617
0.4	35.76	38.69	74.83/74.78	76.08/74.86	1.016	0.576
0.5	33.91	36.88	72.93/72.83	74.07/73.83	1.026	0.524
0.6	29.83	34.08	67.95/67.98	69.77/68.46	1.163	0.444
0.7	27.66	27.45	64.91/64.96	67.10/65.87	1.236	0.379
0.8	25.74	25.80	62.38/62.44	63.43/62.11	1.296	0.326
0.9	23.27	24.34	59.05/59.06	60.43/59.76	1.380	0.244
Avg.	**33.80**	**36.40**	**72.43/72.08**	**73.70/72.76**	**1.049**	**0.528**

Table 15: Robustness evaluation results of DEAR on MOSI under various rates of intra-modal missing data.

Missing Rate r	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
0.0	50.48	52.41	85.34/84.36	85.69/83.76	0.536	0.767
0.1	49.85	51.45	84.31/83.72	84.93/83.37	0.554	0.746
0.2	48.87	50.65	83.19/82.84	83.25/82.20	0.578	0.716
0.3	47.54	48.81	82.27/81.80	82.89/80.50	0.611	0.680
0.4	46.23	47.42	81.43/80.54	80.86/79.65	0.632	0.653
0.5	45.26	45.93	79.58/78.67	79.75/78.57	0.669	0.598
0.6	44.47	45.43	77.87/77.87	76.78/76.82	0.699	0.554
0.7	42.46	43.36	75.46/75.86	74.65/75.83	0.731	0.485
0.8	40.58	41.34	73.15/74.93	71.81/73.26	0.763	0.416
0.9	38.89	39.66	69.95/72.89	69.64/69.51	0.798	0.298
Avg.	**45.46**	**46.65**	**79.26/79.35**	**79.03/78.35**	**0.657**	**0.591**

Table 16: Robustness evaluation results of DEAR on MOSEI under various rates of intra-modal missing data.

Missing Rate r	Acc-5	Acc-3	Acc-2	F1	MAE↓	Corr
0.0	41.35	64.4	79.07	78.67	0.455	0.563
0.1	39.96	62.73	78.34	77.57	0.453	0.559
0.2	39.67	63.09	76.73	74.75	0.462	0.525
0.3	36.91	61.63	76.08	73.79	0.472	0.500
0.4	36.32	59.00	75.42	74.01	0.486	0.491
0.5	35.29	58.49	78.16	71.50	0.502	0.418
0.6	35.04	53.97	72.79	69.19	0.516	0.376
0.7	31.75	54.26	71.05	65.45	0.535	0.329
0.8	24.72	49.96	69.73	61.94	0.567	0.262
0.9	25.38	48.14	69.00	60.99	0.581	0.208
Avg.	**34.64**	**57.57**	**74.64**	**70.79**	**0.503**	**0.423**

Table 17: Robustness evaluation results of DEAR on SIMS under various rates of intra-modal missing data.