ROBUST MULTIMODAL LEARNING VIA DUAL-CONSTRAINED INFORMATIVE LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal systems deployed in real-world scenarios often encounter both missing modalities and corrupted observations. Such inputs make it challenging for existing methods to identify reliable information for effective multimodal interaction. Moreover, noisy and missing modalities are typically handled separately, hindering unified robustness under complex multimodal conditions. To address these challenges, we propose **R**eliability-aware **D**ual-constraint **I**nformation **B**ottleneck (RDIB). RDIB introduces a dual-constraint information bottleneck that filters noisy information while preserving both target-modality and task-relevant information, providing an information-theoretic basis for representation reliability. Building on the informative representations, a unified reliability estimation mechanism combines bottleneck-derived reliability for observed representations with estimated reliability for reconstructed representations. We further construct reliability-aware complete and incomplete fusion streams and impose a fusion-level dual-constraint information bottleneck to learn robust multimodal representations under incomplete inputs. Extensive experiments on MOSEI, D-Vlog, LMVD and Food-101 under diverse modality conditions demonstrate that RDIB consistently achieves robust and competitive performance across multimodal learning scenarios.

1 INTRODUCTION

Multimodal learning (Xu et al., 2023) integrates information from multiple input sources to improve downstream task performance, with various fusion strategies and task-specific architectures developed for applications such as vision-language tasks (Kim & Kim, 2024), action recognition (Woo et al., 2023; Zhou et al., 2022a) and sentiment analysis (Luo et al., 2025; Mai & Han, 2026). However, most existing methods assume that all modalities are available during both training and inference, whereas in practical applications modalities may be missing due to factors such as asynchronous data capture, sensor failure, or privacy concerns. Existing modality-missing approaches, including unified multimodal learning (Hazarika et al., 2020), modality imputation (Yuan et al., 2021), representation alignment (Li et al., 2024) and feature reconstruction (Wang et al., 2023), aim to mitigate the impact of incomplete inputs. More recently, addressing modality-missing problems from an information-theoretic perspective has attracted increasing attention. Several methods adopt information theory to improve multimodal representations under missing modalities (Wu et al., 2025; Lin et al., 2025a; Zhuang et al., 2025; Lin et al., 2025b).

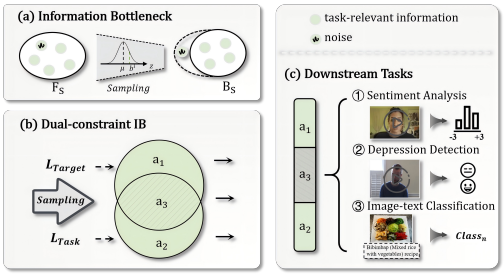


Figure 1: Motivation of DIB. Representation a_1 is mainly optimized toward modality-level information and a_2 emphasizes task-relevant information, while a_3 captures informative space jointly guided by both constraints.

Motivation. Existing multimodal methods based on the information bottleneck generally learn compact representations by suppressing noisy information while retaining information useful for downstream prediction (Xiao et al., 2024; Lin et al., 2025a) as illustrated in Figure 1 (a). How-

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

ever, prior approaches often emphasize one side of this trade-off, such as information compression or task relevance, rather than explicitly characterizing the information jointly constrained by both modality-level targets and downstream objectives. This motivates us to formulate a dual-constraint information bottleneck (DIB), as shown in Figure 1 (b), where the latent representation is jointly optimized by a target-modality objective and a task-relevant objective. The two constraints jointly encourage the bottleneck to preserve target-modality and task-relevant information while penalizing unnecessary information from the source representation. Additionally inspired by HME (Zhuang et al., 2025), which exploits bottleneck uncertainty to characterize modality confidence, we further introduce this intrinsic reliability into multimodal fusion rather than using it only as an auxiliary latent statistic. Based on the dual-constrained latent representations, the resulting reliability-aware fusion produces robust multimodal features for downstream tasks as illustrated in Figure 1 (c). However, extending this mechanism to diverse modality-missing scenarios requires consistent reliability modeling for both observed and reconstructed representations, motivating the development of a unified DIB-based reliability estimator.

Building on these ideas, we introduce **R**eliability-aware **D**ual-constraint **I**nformation **B**ottleneck (RDIB), a unified framework for robust multimodal learning under diverse input conditions. RDIB consists of three key components: ❶ a dual-constraint information bottleneck (DIB), which learns an informative latent representation by suppressing noisy information while preserving target-modal and task-relevant information and simultaneously captures token-wise reliability; ❷ a cross-modal reliability estimator, which provides unified reliability modeling for observed and reconstructed representations and transfers cross-modal information to reconstruct missing-modality features; and ❸ a reliability-aware dual-stream fusion mechanism, which constructs complete and incomplete multimodal streams through reliability-aware multimodal fusion and regularizes the incomplete stream toward the complete multimodal informative space by applying DIB at the fusion level. Through these designs, RDIB maintains robust performance across diverse multimodal input scenarios. Our main contributions can be summarized as follows:

- We propose a Dual-constraint Information Bottleneck (DIB) and provide a theoretical analysis of its information-retention properties.
- Building upon DIB, we develop a Reliability-aware Dual-constraint Information Bottleneck framework for robust multimodal learning, which integrates DIB-based reliability estimation, cross-modal information transfer and reliability-aware multimodal fusion to handle diverse modality conditions.
- Extensive experiments on multiple datasets demonstrate that our method achieves robust and competitive performance compared with existing state-of-the-art multimodal methods.

2 Related Work

2.1 Missing Modality for Multimodal Learning

The missing modality issue refers to the severe performance degradation of multimodal models when specific input streams are unavailable during inference due to sensor failures, privacy constraints, or environmental disruptions (Wu et al., 2024; Lin & Hu, 2023). Existing studies generally address this problem through representation adaptation or modality completion. Alignment-based methods reduce the representation gap between complete and incomplete inputs (Zhang et al., 2020; Pan et al., 2024). Recent work further improves robustness to different missing patterns. SAM-LML (Mai et al., 2025) regards missing modalities as a special case of noisy modalities and introduces supervised attention to jointly suppress noisy information and handle modality missingness. DyMo (Du et al., 2026) formulates missing-modality handling as an inference-time selection problem, adaptively retaining reliable recovered modalities according to task-relevant information. TMDC (Zhuang et al., 2026) jointly considers noisy and missing modalities through sequential intra-modality denoising and inter-modality complementation, improving robustness beyond treating the two degradations independently. Another line reconstructs absent information from observed modalities (Lian et al., 2023). IMDer (Wang et al., 2023) employs a diffusion model conditioned on available modalities to recover missing modality features while maintaining modality-specific distributions and reducing semantic ambiguity. HyperEF (Qiu et al., 2026) employs hypergraph-

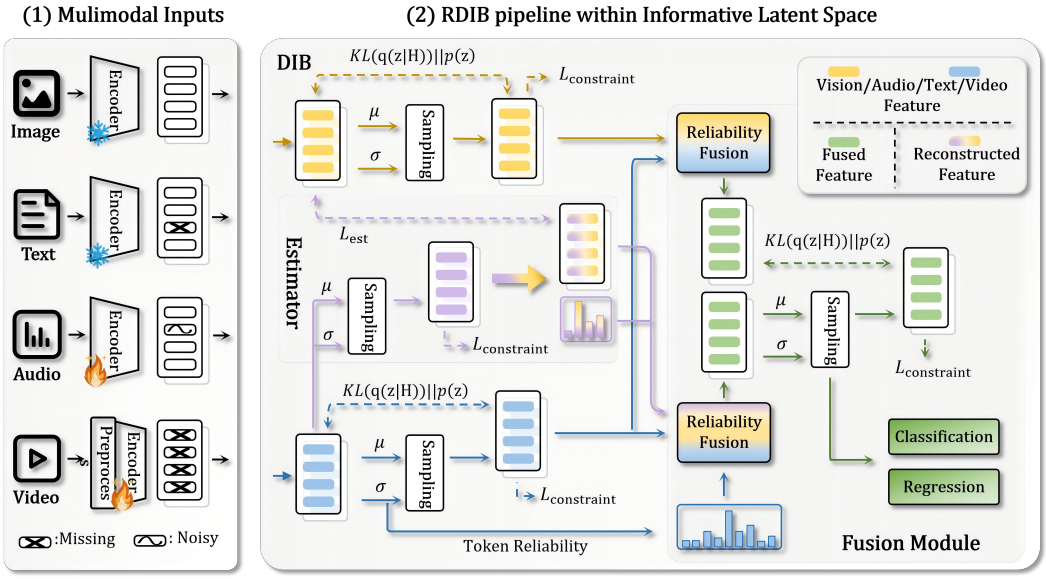


Figure 2: Overview of the RDIB framework. (a) Multimodal inputs are processed by the corresponding preprocessing modules and encoders. (b) The RDIB pipeline within the informative latent space. DIB (Section 3.2) compresses modality representations into an informative latent space and derives token-wise reliability. The Unified Reliability Estimator (Section 3.3) further refines reliability across different modality conditions. Based on the learned representations and reliability, the Reliability-aware Dual-stream Fusion module (Section 3.4) adaptively integrates multimodal information to obtain robust fused representations for downstream tasks.

conditioned diffusion to recover missing latent features and further models feature- and decision-level uncertainty for adaptive evidence fusion.

2.2 Information Bottleneck

The Information Bottleneck (IB) theory (Tishby et al., 2000) provides a principled information-theoretic foundation for representation learning by balancing input compression and the preservation of task-relevant predictive information (Saxe et al., 2019). The Variational Information Bottleneck (VIB) (Alemi et al., 2017) makes this objective tractable through variational approximation. Extending IB to multimodal learning, MIB (Mai et al., 2022) regularizes both unimodal and multimodal representations to suppress modality-specific noise, while ITHP (Xiao et al., 2024) hierarchically compresses multimodal information to construct compact information flows. OMIB (Wu et al., 2025) further derives theoretically grounded regularization constraints and dynamically balances task-relevant information across modalities. Recent studies have extended IB toward robust multimodal learning. CyIN (Lin et al., 2025a) restores missing information in the latent space via cross-modal cyclic translation. IBMA (Wang et al., 2026b) imposes information bottlenecks on both modality-specific and fused representations while enhancing cross-modal semantic alignment through a distribution-free variational objective. For noisy multimodal inputs, SeD-UD (Li et al., 2026) employs input-adaptive bottlenecks and hierarchically decouples redundancy removal from noise suppression. For incomplete multimodal learning, ABDiff (Lin et al., 2025b) incorporates an affection-guided information bottleneck into diffusion-based missing-modality reconstruction to suppress task-unrelated noise and modality-specific redundancy.

3 Methodology

3.1 Task Formulation

Let $\mathcal{M}_d$ denote the set of modalities available in dataset d, where $\mathcal{M}_d = \{X_L, X_V, X_A\}$. In real-world scenarios, some modalities may be missing for individual samples. To simulate

this, missing modality representations are replaced with zero vectors in given datasets $\mathcal{D} = \{(X_L^i, X_V^i, X_A^i, Y^i)\}_{i=1}^N$. Following previous research (Zhuang et al., 2025; Reza et al., 2025; Gao et al., 2026), we employ corresponding preprocessing pipelines to standardize all modalities to the same feature dimension d and sequence length T for different datasets (Preprocessing modules see Appendix A.1), denoted as $\{H_L^i, H_V^i, H_A^i\}_{i=1}^N \in \mathbb{R}^{T\times d}$, respectively. Finally, the objective is to learn a model $\mathcal{M}$ predicting $\hat{y}_i = \mathcal{M}(H_L^i, H_V^i, H_A^i)$ using incomplete modalities. The overall architecture of the proposed method is shown in Figure 2.

3.2 Dual Constraint Information Bottleneck

Given a source representation F_S and a target representation F_T, where $F_S, F_T \in \{H_L, H_V, H_A\}$, DIB encodes F_S into an informative latent representation B_S. Instead of preserving information with respect to only a single target, DIB imposes two complementary constraints on the same bottleneck latent: B_S should retain information relevant to the target modality F_T, while simultaneously preserving information relevant to the downstream label Y. The resulting objective is formulated as:

$$\mathcal{L}_{\text{DIB}} = I(F_S; B_S) - \beta_t I(B_S; F_T) - \beta_y I(B_S; Y), \tag{1}$$

where $I(\cdot;\cdot)$ denotes mutual information, and β_t and β_y control the strengths of the target modality and task constraints, respectively. The two negative mutual information terms constitute the *dual constraint*. DIB seeks a compact latent representation that preserves target-modality information and supports downstream prediction, rather than relying on compression alone.

Variational Information Bottleneck. To optimize Equation 1, we adopt a variational approximation (Alemi et al., 2017) to obtain a tractable training objective. Specifically, the conditional distribution of the bottleneck latent is modeled as

$$q_\theta(B_S|F_S) = \mathcal{N}\left(\mu_S, \text{diag}(\sigma_S^2)\right), \tag{2}$$

with a standard Gaussian prior $r(B_S) = \mathcal{N}(0, \mathbf{I})$. The bottleneck latent can then be sampled using the reparameterization trick,

$$B_S = \mu_S + \sigma_S \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{3}$$

The compression term $I(F_S; B_S)$ admits a variational upper bound, while the two information retention terms admit variational lower bounds. Combining these bounds yields the following tractable surrogate optimization problem $\min_{\theta,\phi,\psi} \mathcal{L}_{\text{dib}}(\theta, \phi, \psi)$:

$$\mathcal{L}_{\text{dib}}(\theta, \phi, \psi) := \mathbb{E}_{p(F_S)}\left[\text{KL}\left(q_\theta(B_S|F_S) \parallel r(B_S)\right)\right] + \beta_t \mathcal{L}_{\text{target}} + \beta_y \mathcal{L}_{\text{task}}. \tag{4}$$

For the two information preservation constraints, the posterior probability $p_\phi(F_T|B_S)$ and $p_\psi(Y|B_S)$ are introduced as variational approximations. The target-modality relevance term $\mathcal{L}_{\text{target}}$ and task relevance term $\mathcal{L}_{\text{task}}$ are formulated as :

$$\mathcal{L}_{\text{target}} = -\mathbb{E}_{B_S\sim q_\theta}\left[\log p_\phi(F_T|B_S)\right] = \mathbb{E}_{B_S\sim q_\theta}\left[\|F_T - D_T(B_S)\|^2\right], \tag{5}$$

$$\mathcal{L}_{\text{task}} = -\mathbb{E}_{B_S\sim q_\theta}\left[\log p_\psi(Y|B_S)\right] = -\mathbb{E}_{B_S\sim q_\theta}\left[\sum_{v=1}^{V} Y_v \log[P_S(B_S)]_v\right], \tag{6}$$

where $D_T(\cdot)$ is used to project bottleneck of the source representation into the target one and $P_S(\cdot)$ is a downstream classification task predictor with V classes. For regression tasks, the posterior probability is computed as $p_\psi(Y|B_S) = e^{-\|Y-\hat{y}_S\|} = e^{-\|Y-P_S(B_S)\|}$.

Different from a conventional information bottleneck that preserves information with respect to a single target (Xiao et al., 2024), DIB constrains the same latent variable from both the modality and task perspectives. The target modal constraint prevents excessive compression from destroying information useful for cross-modal interaction, while the task constraint further distinguishes useful target modal information from task-irrelevant correlations. Their joint optimization guides B_S toward an informative latent that retains information necessary for robust multimodal learning while suppressing unnecessary source-specific content. A detailed derivation of Equation 4 is provided in Appendix B.1.

Reliability Estimation. Inspired by prior work (Zhuang et al., 2025), the posterior uncertainty of an information bottleneck can be used to characterize the reliability of modality representations. In DIB, each source representation F_S is mapped to a Gaussian bottleneck distribution $q_\theta(B_S|F_S) = \mathcal{N}(\mu_S, \mathrm{diag}(\sigma_S^2))$, where μ_S denotes the posterior mean of the latent representation and σ_S^2 denotes its posterior variance. Since the posterior is jointly optimized under the target-modality and task constraints in Equation 4, we use its uncertainty to derive a token-wise reliability score.

Specifically, for the t-th token of modality S, we aggregate the token-wise posterior variance and define its log-precision reliability scores as r_S^t:

$$r_S^t = -\log\left(\frac{1}{D_B}\sum_{d=1}^{D_B}(\sigma_{S,d}^t)^2 + \epsilon\right), \qquad \widetilde{r}_S^t = r_S^t - \frac{1}{T_S}\sum_{j=1}^{T_S} r_S^j. \tag{7}$$

where D_B denotes the bottleneck dimension and ϵ is a constant for numerical stability. We center the log-precision scores to characterize relative reliability within each modality. A larger $\widetilde{r}_S^t$ indicates lower posterior uncertainty and higher reliability relative to other tokens in the same modality.

In this way, DIB produces the informative latent representation and its reliability jointly within the same probabilistic bottleneck.

3.3 Unified Reliability Estimator

We estimate the reliability of reconstructed modalities with respect to the informative target representations learned from observed modalities during training, establishing a consistent reliability interpretation for subsequent multimodal fusion.

Reliability Estimation and Translation for Missing Modalities. For a missing target modality T, we use an available source modality S to recover its informative representation. Instead of directly translating the original source feature, we first construct a cross-modal DIB, $q_\theta^{S\to T}(B_{S\to T}|F_S)$, whose bottleneck latent preserves both target-modality and task-relevant information. We denote the DIB objective imposed on these cross-modal bottlenecks as $\mathcal{L}_{dib}^c$. A probabilistic translator $\mathcal{T}_{S\to T}$ based on CRA (Tran et al., 2017) then maps this latent toward the target bottleneck space and jointly predicts the reconstructed representation $\widehat{B}_{T\leftarrow S}$ and its log-variance $s_{T\leftarrow S}$:

$$(\widehat{B}_{T\leftarrow S}, s_{T\leftarrow S}) = \mathcal{T}_{S\to T}(B_{S\to T}). \tag{8}$$

To jointly optimize cross-modal translation and reliability estimation, we model the target bottleneck representation using a heteroscedastic Gaussian likelihood (Kendall & Gal, 2017),

$$p_\phi(B_T|B_{S\to T}) = \mathcal{N}\left(\widehat{B}_{T\leftarrow S}, e^{s_{T\leftarrow S}}\mathbf{I}\right), \tag{9}$$

where B_T is obtained from the target-modality DIB. Maximizing this likelihood leads to the heteroscedastic negative log-likelihood objective, denoted by estimation objective $\mathcal{L}_{\text{est}}$:

$$\mathcal{L}_{\text{est}}^{S\to T} = \frac{1}{2}\mathbb{E}\left[e^{(-s_{T\leftarrow S})}\mathcal{E}(B_T, \widehat{B}_{T\leftarrow S}) + s_{T\leftarrow S}\right] + \left\|B_T - \widehat{B}_{T\leftarrow S}\right\|^2, \tag{10}$$

The likelihood term jointly learns the conditional mean and uncertainty of the translated target representation: the mean is optimized toward B_T, while the predicted variance reflects the conditional reconstruction uncertainty. The additional MSE term regularizes the reconstruction accuracy.

Consistent with Equation 7, we convert the predicted translation variance into token-wise log-precision reliability:

$$r_{T\leftarrow S}^{est} = -\log\left(e^{s_{T\leftarrow S}^t} + \epsilon\right). \tag{11}$$

A larger $r_{T\leftarrow S}^{est}$ indicates that the target token can be recovered more reliably from the available source modality. The score is further centered within each modality as Equation 7.

Unified Reliability Estimation. Although DIB posterior uncertainty and translation uncertainty arise from different conditional distributions, both uncertainty estimates are associated with the same dual-constrained target representation B_T. When T is observed, its reliability is obtained directly from $q_\theta(B_T|F_T)$. When T is missing, $p_\phi(B_T|B_{S\to T})$ is optimized against the same B_T, and its predicted variance characterizes the conditional uncertainty of recovering this informative target representation. Therefore, $r_{T\leftarrow S}^{\text{est}}$ provides an optimization-consistent surrogate for the unavailable DIB reliability of modality T, allowing the two reliability estimates to be used according to modality availability. A detailed optimization analysis is provided in Appendix B.2.

3.4 RELIABILITY-AWARE DUAL-STREAM FUSION

Building upon the informative latent representations and their reliability, we propose Reliability-aware Dual-stream Fusion (RDF) to perform reliability-aware multimodal interaction and construct the final multimodal representation.

Reliability-aware multimodal fusion. For each modality $m \in \mathcal{M} = \{L, V, A\}$, we denote its DIB latent representation and reliability as B_m and r_m, respectively. When modality m is missing, its reconstructed latent representation and estimated reliability are denoted by $\widehat{B}_m$ and r_m^{est}. Given the modality availability indicator $a_m \in \{0, 1\}$, the representation and reliability actually used for fusion are selected as

$$\widetilde{B}_m = a_m B_m + (1 - a_m)\widehat{B}_m, \qquad r_m^{\text{fus}} = a_m \widetilde{r}_m + (1 - a_m)\widetilde{r}_m^{\text{est}}. \tag{12}$$

We then incorporate the corresponding token-wise reliability into cross-modal attention as an additive prior. For the interaction from modality j to modality i,

$$\mathbf{F}_{\widetilde{B}_i \leftarrow \widetilde{B}_j} = \text{Softmax}\left(\frac{\mathbf{Q}_{\widetilde{B}i}\mathbf{K}_{\widetilde{B}_j}^{\top}}{\sqrt{D}} + r_j^{\text{fus}}\right)\mathbf{V}_{\widetilde{B}j}, \qquad i \neq j, \tag{13}$$

where r_j^{fus} is broadcast along the query dimension and thus adjusts the attention assigned to each token of the source modality j. The cross-modal interaction features are subsequently concatenated and aggregated to obtain

$$F_{\text{est}}^f = \text{ATTN}\left(\text{Concat}\left(\mathbf{F}_{\widetilde{B}_i \leftarrow \widetilde{B}j} i \neq j\right)\right). \tag{14}$$

Accordingly, observed modalities contribute their DIB representations and corresponding reliability, whereas a missing modality is replaced by its reconstructed representation and estimated reliability.

Dual-stream fusion. During training, we construct a complete informative stream from the unmasked modality representations and their DIB-derived reliability scores. Applying the same reliability-aware fusion yields the complete fused representation F^f. In contrast, F_{est}^f is obtained from the availability-dependent representations defined above and therefore represents the information accessible under incomplete multimodal inputs.

Following the DIB formulation in Section 3.2, we take F_{est}^f as the source representation and F^f as the target representation and encode F_{est}^f into the fusion bottleneck B_f. The corresponding DIB objective in token-level form is

$$\mathcal{L}_{dib}^f \approx \frac{1}{T}\sum_{i=1}^{T} KL\left(p_\theta(b_f^i \mid f_{est}^i) \| \mathcal{N}(0, \mathbf{I})\right) + \beta_t \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_{b_f^i}\left[\left|f^i - D_f(b_f^i)\right|^2\right] + \beta_y \mathcal{L}_{task}^f,$$

where f_{est}^i and f^i denote the i-th tokens of the estimated and complete fusion streams, respectively. In this way, the incomplete information stream is compressed toward the complete multimodal informative space under the same dual constraints defined in Section 3.2.

Finally, the fusion bottleneck is aggregated along the temporal dimension and passed to the prediction head: $\hat{y} = C\left(\text{GAP}(B_f)\right)$, where $\text{GAP}(\cdot)$ denotes global average pooling and $C(\cdot)$ denotes the final prediction head.

3.5 TOTAL OPTIMIZATION OBJECTIVE

The proposed framework is jointly optimized by the downstream prediction objective, the DIB objectives and the reliability estimation objective. We introduce hyperparameters λ_{dib} and λ_{est} to balance information bottleneck learning and cross-modal reliability estimation, respectively. The complete objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{task}^{final} + \lambda_{dib}(\mathcal{L}_{dib} + \mathcal{L}_{dib}^f) + \lambda_{est}(\mathcal{L}_{est} + \mathcal{L}_{dib}^c) \tag{15}$$

where $\mathcal{L}_{task}^{final}$ is computed from the final prediction $\hat{y}$ using the standard loss corresponding to the downstream task.

Table 1: Performance comparison under different modality availability settings on D-Vlog and LMVD. The best results are in **bold** and the second are underlined.

Datasets	Availability	TAMFN	LNLN	DepMamba	HME	P^{3}HF	Ours
D-Vlog (F1)	Vision	69.87	68.95	70.89	70.07	70.11	**74.30**
	Audio	71.14	71.82	70.59	72.87	71.32	**74.55**
	Both	73.14	72.98	73.83	73.28	72.49	**74.81**
	Average	71.38	71.25	71.77	72.07	71.31	**74.55**
LMVD (F1)	Vision	69.78	70.67	71.09	71.29	70.20	**73.10**
	Audio	58.50	62.63	**65.54**	62.99	60.61	65.02
	Both	70.13	70.27	73.20	72.39	70.83	**75.27**
	Average	66.14	67.86	69.94	68.89	67.21	**71.13**

4 Experiments

4.1 Experiment Setup

Datasets. ❶ D-Vlog (Yoon et al., 2022) and LMVD (He et al., 2025) are audiovisual datasets for multimodal depression detection. We report Accuracy (ACC) and F1-score (F1) as the metrics. ❷ Food-101 (Wang et al., 2015) is an image–text dataset for food recognition across 101 classes. We adopt Accuracy (ACC) as the metric. ❸ MOSEI (Zadeh et al., 2018) is a tri-modal dataset for multimodal sentiment analysis. We report 'positive/negative' Accuracy (ACC) and F1-score (F1) as the metrics.

Baselines. We compare RDIB with several state-of-the-art methods for different datasets. ❶ TAMFN (Zhou et al., 2022c), LNLN (Zhang et al., 2024), DepMamba (Ye et al., 2025), HME (Zhuang et al., 2025), P^{3}HF (Fu et al., 2026) are used for audiovisual task; ❷ CoOP (Zhou et al., 2022b), MAPs (Lee et al., 2023), DCP (Hu et al., 2024), SyP (Zhang et al., 2025), AOEPT (Lang et al., 2026) are used for the image–text task. ❸ GCNet (Lian et al., 2023), IMDer (Wang et al., 2023), LNLN (Zhang et al., 2024), HME (Zhuang et al., 2025), API (Wang et al., 2026a) are used for MSA task.

Missing Modality Setting. We adopt a rigorous approach wherein modality absence occurs throughout both the training and inference phases. ❶ D-Vlog and LMVD use a fixed missing-modality protocol in which one modality is masked according to the specified missing pattern. ❷ MOSEI uses missing rate (MR) as random missing protocol ranging from 0.1 to 0.7. ❸ Food-101 uses a sample-level missing rate of 70% during both training and inference and considers three missing-modality scenarios: text missing, image missing and both missing.

Due to space limitations, details on datasets description, implementation, baselines and evaluation metrics are provided in Appendix C. In addition, extensive additional experimental results are provided in Appendix D.

4.2 Comparison Analysis

Table 2: Dataset statistics and encoder settings for the evaluated multimodal tasks.

Dataset	Pretrained Encoder	Encoder Training	Avg. Duration
MOSEI	✓	✗	7.28 s
D-Vlog	✗	✓	596.29 s
LMVD	✗	✓	423.82 s
Food-101	✓	✗	-

We present the comparison results in Table 1, Table 3 and Figure 3, covering three multimodal learning settings: regression on MOSEI, long-sequence sample classification on D-Vlog and LMVD and large-scale multi-class classification on Food-101 with 101 categories. Across these datasets and different modality-missing scenarios, RDIB consistently achieves the best averaged performance, demonstrating its robustness to diverse task formulations, sequence characteristics and missing conditions. In particular, compared with HME and LNLN, which are evaluated across MOSEI, D-Vlog and LMVD, RDIB exhibits more consistent performance across regression and long-sequence classification tasks, indicating stronger adaptability to different multimodal learning scenarios.

Table 3: Performance comparison on MOSEI under different missing rates. † denotes results reported in HME (Zhuang et al., 2025).

Datasets	MR	GCNet†	IMDer†	LNLN†	API	HME†	Ours
MOSEI (ACC/F1)	0.1	82.3/82.1	83.5/83.3	83.1/83.0	83.5/83.5	84.0/83.8	**84.2**/83.7
	0.2	80.3/79.9	81.3/80.9	81.7/81.0	81.1/80.9	82.3/81.9	**82.6**/81.1
	0.3	77.5/76.8	79.4/78.8	79.1/78.1	78.4/78.4	80.0/79.6	**80.1**/79.2
	0.4	76.0/74.9	76.1/74.3	76.9/75.6	76.4/75.0	77.6/76.4	77.4/**76.8**
	0.5	74.9/73.2	75.3/72.4	75.1/72.4	74.3/73.2	75.7/74.7	**75.8/75.3**
	0.6	74.1/72.1	71.2/66.5	72.5/68.6	71.6/69.7	73.0/72.5	73.9/**73.3**
	0.7	73.2/70.4	71.1/65.9	70.6/65.2	68.9/66.5	72.4/69.2	72.6/69.1
	Avg.	76.9/75.6	76.8/74.6	77.0/74.8	76.3/75.3	77.9/76.9	**78.1/76.9**

Table 4: Ablation study on LMVD across three missing-modality settings and on MOSEI across three selected missing rates.

Datasets	**Full Avg.**	w/o DIB	w/o L_{target}	w/o L_{task}	w/o Est	w/o DF	w/o RF
LMVD	**71.13**	68.87	69.55	69.71	67.68	69.32	68.55
MOSEI	**78.1/76.5**	76.7/74.7	76.7/74.8	76.8/74.7	76.8/74.2	76.9/74.8	76.5/74.6

4.3 Ablation Study

To better understand the contribution of each component in RDIB, we conduct an ablation study on several key modules of the proposed framework: (i) 'w/o DIB', which removes the dual-constraint information bottleneck; (ii)'w/o L_{target}', which removes the target-modality constraint from DIB; (iii) 'w/o L_{task}', which removes the task-relevant constraint from DIB; (iv) 'w/o Est', which removes the reliability estimator and its corresponding estimation constraint; (v) 'w/o DF', which removes the dual-stream fusion mechanism; (vi) 'w/o RF', which removes reliability-aware multimodal fusion. The average results are reported in Table 4. Full results and additional analyses are provided in Appendix D.2.

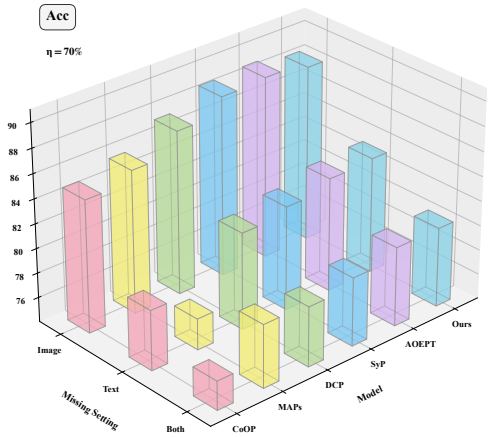


Figure 3: Performance comparison on Food-101 with 70% missing rates across diverse missing scenarios.

Effects of the RDIB Components. Overall, removing most components degrades performance on LMVD and MOSEI, demonstrating the effectiveness of the proposed modules. For the DIB-related components, removing DIB or either of its constraints consistently degrades performance, indicating that target-modality and task-relevant supervision provide complementary guidance for learning informative representations. The performance degradation caused by removing 'Est' or 'RF' highlights the importance of cross-modal reliability estimation, feature reconstruction and reliability-aware fusion. Similarly, removing 'DF' reduces performance, supporting the effectiveness of dual-stream fusion in aligning incomplete representations with the complete informative stream.

4.4 Visual Analysis

To better understand the effect of DIB on representation learning, we visualize the latent representations before and after DIB processing using t-SNE and PCA on LMVD, as shown in Figure 4. DIB largely preserves the original distribution structure while compressing the representations into the latent space, shown in Figure 4 (a) and (b). Meanwhile, the separation between the two classes becomes more pronounced after DIB processing, indicating that DIB suppresses less relevant infor-

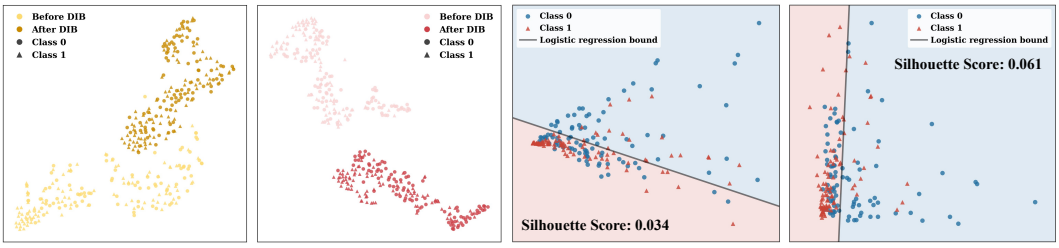


Figure 4: Visualization of latent representations on LMVD before and after DIB processing using t-SNE and PCA.

mation while retaining task-relevant information for downstream prediction. Figure 4 (c) and (d) further present the PCA projections of the two classes, where a logistic regression decision boundary is fitted to provide a view of their separability. To quantitatively evaluate this property, we additionally calculate the Silhouette Score (Rousseeuw, 1987), which measures intra-cluster cohesion relative to inter-cluster separation. The score increases from 0.034 before DIB to 0.061 after DIB, providing quantitative evidence that DIB improves class separability.

4.5 Case Study

To examine how reliability affects multimodal fusion, we randomly select two Food-101 test samples, as shown in Figure 5. We quantify modality-specific attention redistribution using the total variation distance D_m between the projected attention distributions of RDIB and w/o RF, where a larger value indicates a greater change in token-level attention within modality m. Case #1 shows a larger redistribution in the image modality, whereas Case #2 shows a larger change in the text modality. Although both variants correctly classify the samples, RDIB assigns higher probabilities to the ground-truth classes. Figure 6 further shows that RDIB shifts attention toward category-relevant image regions in Case #1 and more informative textual tokens in Case #2, while the image tokens in Case #2 also become more concentrated on category-relevant regions. These observations qualitatively support the role of reliability in guiding multimodal fusion toward task-relevant information.

	Input		Distance	DIB	w/o RF
#1	image 	text: apple pie and cake recipes -health.com	$D_v = 0.422$ $D_a = 0.210$	0.9999	0.9998
#2	image 	text: cinnamon ice cream recipe - cooking | add a pinch | robyn stone	$D_v = 0.274$ $D_a = 0.290$	0.9999	0.9989

Figure 5: Case study of two Food-101 test samples, including attention variation distance D_m and prediction confidence produced by RDIB and without reliability-aware fusion (w/o RF).

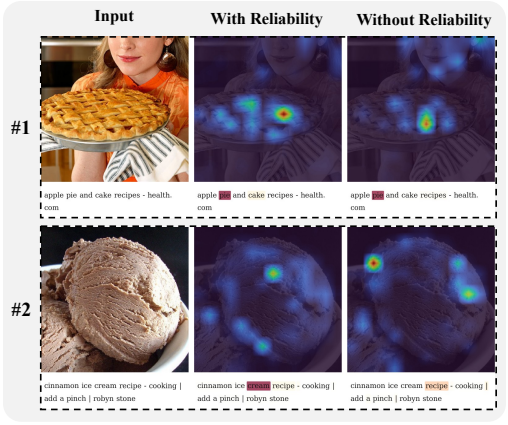


Figure 6: Visualization of attention weights for two Food-101 test samples.

5 Conclusion

In this paper, we propose Reliability-aware Dual-constraint Information Bottleneck (RDIB), a unified framework for robust multimodal learning across diverse missing-modality settings. RDIB learns robust multimodal representations by jointly preserving target-modality and task-relevant information with a dual-constraint information bottleneck, estimating modality reliability through a unified estimator and performing reliability-aware dual-stream fusion for downstream prediction.

AI USE STATEMENT

Large language models (LLMs) were used as writing assistance tools during manuscript preparation and also used to review the completeness and logical consistency of the theoretical analysis, while all theoretical derivations and verification were independently conducted by the authors. LLMs did not contribute to the core research ideas, experimental design or technical innovations. The authors take full responsibility for the accuracy, validity and integrity of all scientific claims and experimental results.

ETHICS STATEMENT

This work uses publicly available benchmark datasets (MOSEI, D-Vlog, LMVD and Food-101). Specifically, D-Vlog and LMVD involve sensitive mental health information for depression detection. We strictly adhere to the ethical guidelines and data usage requirements established by the original dataset authors. For both datasets, we exclusively use the provided pre-extracted features without accessing raw audiovisual recordings or personally identifiable information. No additional human data is collected and all experimental results are reported in aggregate to protect participant privacy. Our method is intended solely for research purposes and should not be used for clinical diagnosis or individual profiling without appropriate validation and ethical oversight.

REPRODUCIBILITY STATEMENT

The model architecture and experimental setup are described in Sections 3 and 4, with implementation details and hyperparameters in Appendix C. Complete theoretical derivations are provided in Appendices B.1 and B.2. Data preprocessing is detailed in Appendix A.1.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (NAACL)*, 2019.

Siyi Du, Xinzhe Luo, Declan O'regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations*, 2026.

Changzeng Fu, Shiwen Zhao, Yunze Zhang, Zhongquan Jian, Shiqi Zhao, and Chaoran Liu. Personality-guided public-private domain disentangled hypergraph-former network for multimodal depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Manning Gao, Tingyi Liu, Leheng Zhang, Haifeng Hu, Yuncheng Jiang, and Sijie Mai. Uncovering latent depression severity for binary depression detection via advantage-weighting ranking. *arXiv preprint arXiv:2607.05901*, 2026.

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

Lang He, Kai Chen, Junnan Zhao, Yimeng Wang, Ercheng Pei, Haifeng Chen, Jiewei Jiang, Shiqing Zhang, Jie Zhang, Zhongmin Wang, et al. Lmvd: A large-scale multimodal vlog dataset for depression detection in the wild. *Information Fusion*, pp. 103632, 2025.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. *Advances in Neural Information Processing Systems*, 37:67446–67466, 2024.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pp. 171–187. Springer, 2024.

Jian Lang, Rongpei Hong, Ting Zhong, and Fan Zhou. Aoept: Breaking the implicit modality-reduction bottleneck in modality-missing prompt tuning. In *International Conference on Machine Learning (ICML)*, 2026.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

Qin Li, Wenbo Zhang, Limei Liu, Han Peng, Junfeng Yang, and Guanying Xu. Sed-ud: An influence-driven and hierarchically-decoupled information bottleneck for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023.

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

Ronghao Lin, Qiaolin He, Sijie Mai, Ying Zeng, Aolin Xiong, Li Huang, Yap-Peng Tan, and Haifeng Hu. Cyin: Cyclic informative latent space for bridging complete and incomplete multimodal learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025a.

Ronghao Lin, Qiaolin He, Ying Zeng, Sijie Mai, Li Huang, and Haifeng Hu. Affection-guided bottleneck diffusion for missing modality issue in multimodal affective computing. *IEEE Transactions on Affective Computing*, 2025b.

Miaosen Luo, Yuncheng Jiang, and Sijie Mai. Towards explainable fusion and balanced learning in multimodal sentiment analysis. In *Proceedings of the 33rd ACM international conference on multimedia*, pp. 1997–2006, 2025.

Sijie Mai and Shiqin Han. Learning invariant modality representation for robust multimodal learning from a causal inference perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45698–45721, 2026.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2022.

Sijie Mai, Shiqin Han, and Haifeng Hu. Supervised attention mechanism for low-quality multimodal data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21377–21397, 2025.

Yuchen Pan, Junjun Jiang, Kui Jiang, and Xianming Liu. Disentangled-multimodal privileged knowledge distillation for depression recognition with incomplete multimodal data. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 5712–5721, 2024.

Xihang Qiu, Yuhao Fang, Qing Zhou, Bin Zhai, Jialong Hong, Wanpeng Zhang, Yao Lu, Ye Zhang, and Chun Li. Beyond missing modalities: Hypergraph conditioned diffusion for uncertainty-aware multimodal emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, 2021.

Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and Salman Asif. Robust multimodal learning via cross-modal proxy tokens. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Xiaotao Wang, Yiyang Fang, Wenke Huang, Bin Yang, Guancheng Wan, and Mang Ye. Api: Adaptive prototype imputation for incomplete multimodal sentiment analysis. In *Forty-third International Conference on Machine Learning (ICML)*, 2026a.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International conference on multimedia & expo workshops (ICMEW)*, pp. 1–6. IEEE, 2015.

Yancheng Wang, Zeyu Dong, Dongfang Sun, Alvin C Silva, Teresa Wu, and Yingzhen Yang. IBMA: Information bottleneck-based multimodal alignment. In *Forty-third International Conference on Machine Learning*, 2026b.

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

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2776–2784, 2023.

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *Forty-second International Conference on Machine Learning (ICML)*, 2025.

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

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12113–12132, 2023.

Jiaxin Ye, Junping Zhang, and Hongming Shan. Depmamba: Progressive fusion mamba for multimodal depression detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12226–12234, 2022.

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

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics*, 2018.

Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 44 (5):2402–2415, 2020.

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

Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, and Rong Jin. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Li Zhou, Zhenyu Liu, Zixuan Shangguan, Xiaoyan Yuan, Yutong Li, and Bin Hu. Tamfn: Time-aware attention multimodal fusion network for depression detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:669–679, 2022c.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hypermodality enhancement for multimodal sentiment analysis with missing modalities. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. Tmdc: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

APPENDIX

A Methodology Details **15**

A.1 Preprocessing Details 15

A.2 Generalization to Multiple Available Modalities 15

B Methodological Analysis **16**

B.1 Derivations of the dual-constraint information bottleneck 16

B.2 Derivations of the reliability for missing modalities 18

C Further Experimental Details **19**

C.1 Dataset Description 19

C.2 Implementation Details 19

C.3 Baseline Methods 19

C.4 Evaluation Metrics 21

C.5 Missing Protocols 21

D Additional Experimental Results **21**

D.1 Results under Noisy Modalities 21

D.2 Detailed Ablation Study 22

D.3 Parameter Sensitivity Analysis 22

D.4 Model Complexity Analysis 23

D.5 Further Analysis 24

A METHODOLOGY DETAILS

A.1 PREPROCESSING DETAILS

For Food-101, following previous work (Reza et al., 2025), we use the pretrained CLIP ViT-B/16 (Radford et al., 2021) to extract visual and textual features, with ViT-B serving as the visual encoder and the CLIP text Transformer as the text encoder. The extracted features are further processed by a semantic proxy tokenizer to obtain representations with a unified token length and hidden dimension.

For D-Vlog and LMVD, following previous work (Gao et al., 2026), we directly use the provided pre-extracted features. Each feature sequence is linearly interpolated to a fixed temporal length T, followed by a 1D convolutional projection to map modality-specific feature dimensions into a shared hidden dimension D_h. The projected features are then encoded using Sequential Time Delay Neural Network (Seq-TDNN) (Gao et al., 2026).

For MOSEI, following previous work (Zhuang et al., 2025), textual features are extracted using pre-trained BERT-base (Devlin et al., 2019). Visual features are obtained using MTCNN and OpenFace, while acoustic features are extracted using COVAREP. For all modalities, a 1D temporal convolutional layer is applied to project the extracted features into the shared hidden dimension D_h.

A.2 GENERALIZATION TO MULTIPLE AVAILABLE MODALITIES

The cross-modal translation in Section 3.3 is formulated for a single source-target pair. When multiple modalities remain available, each source can provide a reconstruction of the same missing modality. To accommodate different availability patterns, we aggregate these reconstructions in the target informative space according to their estimated reliability. Let $\mathcal{A} = \{S \in \mathcal{M} \mid a_S = 1\}$

denote the nonempty set of observed modalities for a sample. For a missing target $T \notin \mathcal{A}$, each source $S \in \mathcal{A}$ produces a candidate $\widehat{B}_{T\leftarrow S}$ and its log-variance $s_{T\leftarrow S}$ through the corresponding cross-modal DIB and estimator.

Specifically, let $\mathcal{A}^t \subseteq \mathcal{A}$ contain the sources with a valid token at position t, and write $r^{\mathrm{est},t}_{T\leftarrow S} = -\log(e^{s^t_{T\leftarrow S}} + \epsilon)$ for the token-wise reliability in Equation 11. For $\mathcal{A}^t \neq \varnothing$, the aggregated reconstruction is computed as:

$$\alpha^t_{T\leftarrow S} = \frac{\exp(r^{\mathrm{est},t}_{T\leftarrow S})}{\sum_{U\in\mathcal{A}^t}\exp(r^{\mathrm{est},t}_{T\leftarrow U})}, \qquad \widehat{B}^t_T = \sum_{S\in\mathcal{A}^t} \alpha^t_{T\leftarrow S}\widehat{B}^t_{T\leftarrow S}. \tag{16}$$

Here the weights are normalized over available sources at each token, assigning larger contributions to candidates with lower predicted uncertainty. We use the reliability scores before temporal centering to preserve uncertainty differences across sources. Unavailable sources and padded tokens receive zero weight.

We propagate reconstruction uncertainty by matching the moments of the mixture of candidate translation distributions with weights $\alpha^t_{T\leftarrow S}$. Averaging the resulting variance over the D_B latent dimensions gives:

$$u^t_T = \sum_{S\in\mathcal{A}^t} \alpha^t_{T\leftarrow S}\left[e^{s^t_{T\leftarrow S}} + \frac{1}{D_B}\left\|\widehat{B}^t_{T\leftarrow S} - \widehat{B}^t_T\right\|_2^2\right], \qquad r^{\mathrm{est},t}_T = -\log(u^t_T + \epsilon). \tag{17}$$

The resulting scores are then centered over valid tokens as in Equation 7, providing $\widetilde{r}^{\mathrm{est}}_T$ together with $\widehat{B}_T$ for the incomplete fusion stream in Section 3.4.

Taking a missing visual modality as an example, the language and acoustic modalities provide $\widehat{B}_{V\leftarrow L}$ and $\widehat{B}_{V\leftarrow A}$, which are combined using their normalized reliability weights. When only one source is available, its weight becomes one and the rule reduces to the single-source reconstruction and uncertainty. A token with no valid source is masked during fusion. Observed modalities retain their own DIB representations and reliability. If several modalities are missing, each is reconstructed directly from the observed sources; reconstructed features are not reused as translation inputs. Thus, the same pairwise translators support different availability patterns without additional translators for specific source combinations.

B Methodological Analysis

B.1 Derivations of the dual-constraint information bottleneck

All information quantities below are evaluated under $p(F_S, F_T, Y)q_\theta(B_S|F_S)$, with a fixed distribution of preprocessed features and nonnegative β_t, β_y. Thus $B_S \perp (F_T, Y) \mid F_S$; no conditional independence between F_T and Y is assumed. We assume that the information quantities, entropies, and expected losses used below are finite, and write $R = I(F_S; B_S)$ for brevity.

Feasibility and the information required by both constraints. For prescribed retention levels $a_t, a_y \geq 0$, the constrained form associated with Equation 1 is

$$\inf_{q_\theta} R \quad \text{subject to} \quad I(B_S; F_T) \geq a_t, \qquad I(B_S; Y) \geq a_y. \tag{18}$$

Its Lagrangian is Equation 1 plus the encoder-independent constant $\beta_t a_t + \beta_y a_y$. Feasibility requires $a_t \leq I(F_S; F_T)$ and $a_y \leq I(F_S; Y)$ by data processing. A sufficient construction is available when finite-rate encoders B_t and B_y satisfy the respective constraints and the latent dimension accommodates their concatenation. Sampling them independently given F_S and setting $B_S = (B_t, B_y)$ preserves both constraints, with $I(F_S; B_t, B_y) \leq I(F_S; B_t) + I(F_S; B_y)$. This follows from the chain rule and $I(B_t; B_y|F_S) = 0$; concatenating diagonal Gaussian encoders also remains within the posterior family in Section 3.2.

The minimum information required by the two constraints depends on their overlap. The chain rule gives the exact identity

$$\begin{aligned} R &= I(B_S; F_T, Y) + I(F_S; B_S|F_T, Y) \\ &= I(B_S; F_T) + I(B_S; Y) - I(F_T; Y) + I(F_T; Y|B_S) + I(F_S; B_S|F_T, Y). \end{aligned} \tag{19}$$

The first equality uses $I(B_S; F_T, Y|F_S) = 0$, and the second expands $I(B_S; F_T, Y)$ in both orders. Nonnegativity of the conditional terms, together with $R \geq I(B_S; F_T), I(B_S; Y)$, yields

$$R \geq \max\{a_t,\ a_y,\ a_t + a_y - I(F_T; Y)\}. \tag{20}$$

Consequently, adding label retention to target-only retention raises this necessary lower bound whenever the new requirement exceeds the information already accounted for. If F_T and Y are independent, the bound becomes $a_t + a_y$; correlated supervision can share the retention budget. Any positive retention level excludes an input-independent bottleneck, although Equation 20 alone does not establish feasibility at a prescribed latent dimension.

How the two constraints select information. Consider an augmented encoder $B'_S = (B_S, Z)$, where Z is additional information encoded from F_S. Applying the chain rule to Equation 1 gives

$$\Delta\mathcal{L}_{\mathrm{DIB}} = I(F_S; Z|B_S) - \beta_t I(Z; F_T|B_S) - \beta_y I(Z; Y|B_S). \tag{21}$$

Retaining Z improves the objective exactly when its weighted target and label gains exceed its compression cost. The label term can therefore justify retaining predictive information unrewarded by target-only supervision, while the target term can preserve information useful for cross-modal decoding beyond label prediction. Information with neither conditional gain only increases the objective if its encoding cost is positive. Conversely, a target-related but label-irrelevant component can still be retained when its target gain outweighs its cost; noise suppression depends on this trade-off.

The weights also determine whether nontrivial retention is preferable to collapse. Data processing implies $\mathcal{L}_{\mathrm{DIB}} \geq (1 - \beta_t - \beta_y)R$. Hence an input-independent encoder is a global minimizer of the information objective when $\beta_t + \beta_y \leq 1$. To strictly improve upon it, an encoder must satisfy $\beta_t I(B_S; F_T) + \beta_y I(B_S; Y) > R$. Thus $\beta_t + \beta_y > 1$ is necessary, but is not by itself sufficient, for a strict improvement over collapse. This conclusion concerns the standalone information objective in Equation 1 under a fixed feature distribution, rather than the joint training objective in Equation 15. At the boundary $\beta_t + \beta_y = 1$, it establishes that an input-independent encoder is a global minimizer, without asserting that all global minimizers are input-independent.

In practice, RDIB optimizes the joint empirical objective in Equation 15, which combines variational DIB losses with separately weighted final-prediction supervision. Therefore, the best observed D-Vlog setting $\beta_t = \beta_y = 0.5$ does not contradict the above result for the standalone information objective.

Variational bounds and their joint approximation gap. Let $q_\theta(B_S)$ be the aggregate posterior, and let $q_\theta(F_T|B_S)$ and $q_\theta(Y|B_S)$ be the conditionals induced by the above joint distribution. Define

$$\begin{aligned} \mathcal{K} &= \mathbb{E}_{p(F_S)}\mathrm{KL}\big(q_\theta(B_S|F_S) \parallel r(B_S)\big), & \delta_0 &= \mathrm{KL}\big(q_\theta(B_S) \parallel r(B_S)\big), \\ \delta_t &= \mathbb{E}_{q_\theta(B_S)}\mathrm{KL}\big(q_\theta(F_T|B_S) \parallel p_\phi(F_T|B_S)\big), & \delta_y &= \mathbb{E}_{q_\theta(B_S)}\mathrm{KL}\big(q_\theta(Y|B_S) \parallel p_\psi(Y|B_S)\big). \end{aligned} \tag{22}$$

Writing $\mathcal{L}_{\mathrm{target}}$ and $\mathcal{L}_{\mathrm{task}}$ as the expected negative log-likelihoods of normalized decoders, direct expansion of these KL divergences yields

$$\begin{aligned} &\mathcal{K} = R + \delta_0, \\ &I(B_S; F_T) = H(F_T) - \mathcal{L}_{\mathrm{target}} + \delta_t, \\ &I(B_S; Y) = H(Y) - \mathcal{L}_{\mathrm{task}} + \delta_y. \end{aligned} \tag{23}$$

These identities give a compression upper bound and two retention lower bounds. In particular, setting $\ell_t = \max\{0, H(F_T) - \mathcal{L}_{\mathrm{target}}\}$ and $\ell_y = \max\{0, H(Y) - \mathcal{L}_{\mathrm{task}}\}$ gives the combined interval

$$\max\{\ell_t, \ell_y, \ell_t + \ell_y - I(F_T; Y)\} \leq R \leq \mathcal{K}. \tag{24}$$

At fixed feature marginals and the other prediction loss, reducing either loss cannot decrease the left endpoint. The right endpoint controls the total encoding budget; the two constraints impose no additional prior-matching term.

Let $\widetilde{\mathcal{L}}_{\mathrm{DIB}}$ denote the right-hand side of Equation 4, interpreted with these normalized likelihoods, and put $C_{\mathrm{ent}} = \beta_t H(F_T) + \beta_y H(Y)$. Substitution of Equation 23 gives

$$\widetilde{\mathcal{L}}_{\mathrm{DIB}} - C_{\mathrm{ent}} - \mathcal{L}_{\mathrm{DIB}} = \delta_0 + \beta_t\delta_t + \beta_y\delta_y \geq 0. \tag{25}$$

Thus the tractable objective, after the entropy shift, upper-bounds Equation 1. Equality requires the aggregate posterior to match the prior and every positively weighted decoder to match its induced conditional almost everywhere. For fixed distributions and decoders, increasing either weight amplifies its approximation gap; stronger retention does not automatically tighten the upper bound. Minimizing the surrogate is therefore justified as upper-bound optimization, rather than a guarantee of attaining the global information optimum. The constant-shift equivalence holds for fixed feature marginals; it does not discard changes in $H(F_T)$ if the target feature extractor is also updated.

Resulting optimization objective. Taking $p_\phi(F_T|B_S) = \mathcal{N}(D_T(B_S), \frac{1}{2}\mathbf{I})$ makes its negative log-likelihood equal to the squared reconstruction error plus a fixed normalization constant. Taking $p_\psi(Y|B_S)$ to be categorical with probabilities $P_S(B_S)$ gives the classification cross-entropy. Removing constants and replacing the data expectation by the training average yields

$$\min_{\theta,\phi,\psi} \frac{1}{N}\sum_{i=1}^{N}\Big[\mathrm{KL}\big(q_\theta(B_S|F_S^i) \parallel r(B_S)\big) + \mathbb{E}_\epsilon\Big[\beta_t\|F_T^i - D_T(B_S^i)\|^2 - \beta_y \sum_{v=1}^{V} Y_v^i \log[P_S(B_S^i)]_v\Big]\Big], \tag{26}$$

where $B_S^i = \mu_S^i + \sigma_S^i \odot \epsilon$ uses the sampling rule in Section 3.2. Other fixed isotropic target variances rescale the reconstruction coefficient and can be absorbed into β_t. This is the empirical form of Equation 4; Equation 25 identifies precisely how its two preservation losses approximate the original dual constraint objective.

B.2 Derivations of the reliability for missing modalities

Conditional uncertainty of the estimate. Let ℓ_{est}^t denote the token contribution to Equation 10, and define its conditional reconstruction error by

$$R_t = \frac{1}{D_B}\mathbb{E}\Big[\|B_T^t - \widehat{B}_{T\leftarrow S}^t\|^2 \mid B_{S\to T}\Big]. \tag{27}$$

Its conditional expected loss is $(\frac{1}{2}e^{-s_{T\leftarrow S}^t} + D_B)R_t + \frac{1}{2}s_{T\leftarrow S}^t$. Differentiating with respect to the log-variance, with the prediction held fixed, gives

$$\frac{\partial \mathbb{E}[\ell_{\text{est}}^t \mid B_{S\to T}]}{\partial s_{T\leftarrow S}^t} = \frac{1}{2}\left(1 - e^{-s_{T\leftarrow S}^t}R_t\right). \tag{28}$$

For $R_t > 0$, the unique minimum satisfies $e^{s_{T\leftarrow S}^{t,*}} = R_t$. Thus underestimating the residual increases the learned variance, while overestimating it decreases the variance. Since the coefficient of R_t is positive, the optimal prediction is $\widehat{B}_{T\leftarrow S}^{t,*} = \mathbb{E}[B_T^t \mid B_{S\to T}] = \mathbb{E}[\mu_T^t \mid B_{S\to T}]$; the additional MSE preserves this optimum.

The relationship to native DIB uncertainty follows directly from conditional variance decomposition:

$$e^{s_{T\leftarrow S}^{t,*}} = \mathbb{E}[v_T^t \mid B_{S\to T}] + \frac{1}{D_B}\mathbb{E}\big[\|\mu_T^t - \mathbb{E}[\mu_T^t \mid B_{S\to T}]\|^2 \mid B_{S\to T}\big]. \tag{29}$$

To obtain this identity, substitute the target reparameterization into R_t: the cross term vanishes because the target sampling noise is independent and zero-mean, leaving $\mathbb{E}[v_T^t \mid B_{S\to T}] + D_B^{-1}\mathbb{E}[\|\mu_T^t - \widehat{B}_{T\leftarrow S}^t\|^2 \mid B_{S\to T}]$. At the optimal mean, this yields Equation 29. An imperfect mean adds its squared conditional bias divided by D_B. Hence the estimator inherits DIB posterior uncertainty and additionally measures target information unrecoverable from the source.

Joint optimization and substitution. The same decomposition explains the coupling during training. Holding the translator outputs and target mean fixed, integration over target posterior sampling yields the local derivative

$$\frac{\partial \mathbb{E}_{q_\theta(B_T|F_T)}[\ell_{\text{est}}^t]}{\partial(\sigma_{T,d}^t)^2} = \frac{e^{-s_{T\leftarrow S}^t}}{2D_B} + 1. \tag{30}$$

With gradients propagated through B_T, the estimation loss therefore updates both the translator and the target posterior underlying native reliability. These gradients combine with DIB's compression and dual preservation constraints in Equation 15. Both reliability branches also receive supervision through their use in task prediction and fusion. They are thus jointly trained for reliable use of the same informative target space. Equations 28–29 describe the estimation loss's conditional optimum; additional task and fusion gradients can shift this optimum during joint training.

For an observed target, $v_T^t = D_B^{-1}\mathbb{E}[\|B_T^t - \mu_T^t\|^2 \mid F_T]$ measures uncertainty around its posterior mean. Equation 29 extends this error interpretation to source-conditioned recovery. Applying the same log-precision mapping allows both uncertainty estimates to serve as reliability-related attention priors. Token-wise centering preserves their relative attention weights. If μ_T^t and v_T^t are determined by $B_{S\to T}$, the recovery term vanishes and the two scores coincide. Otherwise, the estimated score includes the additional recovery uncertainty. Provided the source-conditioned target distribution transfers from training to inference, this supports using $r_{T\leftarrow S}^{\text{est}}$ in place of unavailable native reliability for the reconstructed token.

C Further Experimental Details

C.1 Dataset Description

D-Vlog and LMVD (Yoon et al., 2022; He et al., 2025) are collected from user-generated vlogs on public social media platforms, providing naturalistic audiovisual behavioral observations for depression detection. D-Vlog was collected from YouTube and LMVD was collected from four major multimedia platforms: Sina Weibo, Bilibili, TikTok, and YouTube.

Both datasets provide pre-extracted visual and acoustic features. For D-Vlog, the visual modality consists of 68 facial landmarks extracted by dlib, while the acoustic modality is represented by 25 low-level descriptors extracted using OpenSMILE. LMVD provides facial landmark coordinates extracted by dlib. Although LMVD also includes OpenFace-based attributes, such as facial action units, eye gaze, and head pose, we use only facial landmarks to ensure feature consistency across datasets. For the acoustic modality, LMVD provides VGGish features.

MOSEI (Zadeh et al., 2018) is one of the largest multimodal sentiment analysis datasets, comprising more than 22,000 video segments from over 1,000 YouTube speakers across approximately 250 topics. The preprocessing of MOSEI follows prior work (Zhuang et al., 2025).

Food-101 (Wang et al., 2015) is a multimodal classification dataset consisting of 101 food classes. It contains 90,704 image-text pairs, divided into a training set of 67,988 pairs and a test set of 22,716 pairs (Reza et al., 2025).

C.2 Implementation Details

We implement RDIB using the PyTorch framework and conduct all experiments on a single NVIDIA GeForce RTX 3090 GPU, using CUDA 12.8 and PyTorch 2.8.0. We train the model using the AdamW optimizer. Table 5 summarizes the main configurations used in our proposed method. Max-seq denotes the maximum number of tokens per modality, while Shared-dim specifies the unified feature dimension across modalities. Dim of Latent represents the output dimension of DIB, and Ratios of CRA layer indicates the relative number of CRA layers assigned to different modalities. The DIB and estimator loss weights correspond to λ_{dib} and λ_{est}, respectively.

C.3 Baseline Methods

All baselines are briefly described as follows. The symbol $^\heartsuit$ indicates baselines reproduced using publicly available code and the hyperparameter settings reported in original papers. For fair comparison, all baselines follow consistent data preprocessing, missing-modality protocols, training configurations, and evaluation metrics on each dataset.

TAMFN $^\heartsuit$ (Zhou et al., 2022c) constructs a time-aware attention network that combines global temporal convolutional networks, inter-modal feature extraction, and time-aware attention modules to effectively model non-verbal behaviors.

Table 5: The main hyperparameter settings in our proposed method.

Setting	D-Vlog	LMVD	Food-101	MOSEI
Training Epochs	35	40	15	100
Batch Size	32	32	64	32
Learning rate	5e-4	1e-5	1e-3	2e-5
Weight Decay	0.01	0.01	2e-2	5e-4
Max-seq	512	512	77	50
Shared-dim	128	128	256	64
Dim of Latent	128	128	128	256
Ratios of CRA layer	2:1	1:4	2:3	2:2:1
Dimension of RA	[128,64,32]	[128,64,32]	[128,64,32]	[128,64,32]
Layers of Attention	8	2	2	2
Heads of Attention	3	4	4	16
DIB loss weight	0.5	0.1	0.05	0.01
Estimator loss weight	0.005	0.05	0.01	0.003

DepMamba $^{\heartsuit}$ (Ye et al., 2025) presents an audio-visual progressive fusion approach that combines hierarchical contextual modeling with a multimodal collaborative state space model to integrate both cross- and intra-modal information.

P^{3}HF $^{\heartsuit}$ (Fu et al., 2026) adopts a personality-guided public-private domain disentangled Hypergraph-Former network, integrating large language model-driven individual representations and contrastive learning to handle individual differences across diverse behavioral contexts.

GCNet (Lian et al., 2023) employs speaker and temporal graph neural networks to capture conversational dependencies, jointly optimizing classification and reconstruction to handle missing modalities.

IMDer (Wang et al., 2023) employs a diffusion model conditioned on available modalities to recover missing modality features while maintaining modality-specific distributions and reducing semantic ambiguity.

LNLN $^{\heartsuit}$ (Zhang et al., 2024) enhances robustness under incomplete data by learning reliable multimodal representations from available modalities while reducing the negative influence of missing modality information.

HME $^{\heartsuit}$ (Zhuang et al., 2025) addresses missing-modality multimodal learning by constructing enhanced hyper-modality representations to compensate for unavailable modalities and strengthen cross-modal semantic interaction.

API $^{\heartsuit}$ (Wang et al., 2026a) imputes missing modalities using semantic-anchored class-temporal prototypes and instance-adaptive affine modulation to generate sample-specific representations.

CoOP (Zhou et al., 2022b) adapts pretrained vision-language models by replacing hand-crafted prompt contexts with learnable continuous context vectors while keeping the pretrained model frozen.

MAPs (Lee et al., 2023) introduces modality-missing-aware learnable prompts to adapt pretrained multimodal Transformers to different missing-modality patterns in a parameter-efficient manner.

DCP (Hu et al., 2024) designs correlated, input-dependent, and modality-common prompts to exploit cross-layer, input-prompt, and cross-modal correlations for visual recognition with missing modalities.

SyP (Zhang et al., 2025) combines static prompts with dynamically generated prompts through a Dynamic Adapter to adapt multimodal representations to different missing-modality conditions.

AOEPT $^{\heartsuit}$ (Lang et al., 2026) introduces modal-contextualized prompts that encode global modality priors and adapt them to observed inputs to compensate for information from missing modalities.

C.4 Evaluation Metrics

Following prior research (Zhuang et al., 2025; Lang et al., 2026; Ye et al., 2025), we adopt a standard set of classification metrics for datasets: Accuracy (ACC), and F1-score (F1), as defined below:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{31}$$

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{32}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

C.5 Missing Protocols

To evaluate the robustness of RDIB under diverse modality-missing conditions, we adopt three evaluation protocols across different datasets. ❶ **Fixed Missing.** For D-Vlog and LMVD, we adopt a fixed missing protocol, where one modality is masked according to a predefined missing pattern while the remaining modalities are available. ❷ **Random Missing.** For MOSEI, following (Zhuang et al., 2025), we randomly mask modalities for each sample while ensuring that at least one modality remains available. The missing rate (MR) is defined as

$$MR = 1 - \frac{\sum_{i=1}^{N} m_i}{N \times M}, \tag{33}$$

where m_i denotes the number of available modalities in the i-th sample, N is the total number of samples and M is the number of modalities. Since at least one modality must remain available, the theoretical maximum MR is $\frac{2}{3}$ for three modalities. Thus, the nominal setting of $MR = 0.7$ approximates this upper bound, with exactly one modality retained per sample. The same MR is maintained across training, validation, and testing. ❸ **Proportional Missing.** For Food-101, following (Lang et al., 2026), we consider three missing patterns with a missing rate of $\eta\%$: (1) *Text Missing*, where $\eta\%$ of samples contain only images; (2) *Image Missing*, where $\eta\%$ of samples contain only text; and (3) *Both Missing*, where $(\eta/2)\%$ of samples are image-only and another $(\eta/2)\%$ are text-only. The remaining samples retain both modalities in all three settings. We apply these protocols during both training and evaluation, with $\eta = 70$ for the main experiments.

D Additional Experimental Results

D.1 Results under Noisy Modalities

To further evaluate model robustness under complex environments involving both modality incompleteness and noisy observations, we conduct noise robustness experiments on three datasets: D-Vlog, LMVD, and MOSEI.

Noise setting. We introduce Gaussian noise into continuous modality features to evaluate model robustness under varying corruption levels. Noise injection is triggered independently for each sample with a probability of 0.5. Once triggered, Gaussian perturbations are applied to valid feature elements, with the noise scale determined by the standard deviation of valid features within each sample multiplied by $\eta/10$, where η denotes the noise intensity. We evaluate eight intensity levels, ranging from 0.0 to 7.0, representing progressively increasing corruption severity.

Results under Gaussian noise. Figure 7 presents the robustness comparison across different Gaussian noise intensities on D-Vlog, LMVD, and MOSEI. Overall, RDIB consistently outperforms HME across all evaluated noise levels, demonstrating its robustness to noisy multimodal inputs. On D-Vlog, RDIB exhibits a smaller performance degradation as noise intensity increases, maintaining a clear advantage under severe perturbations. On LMVD, both methods experience more pronounced performance degradation, while RDIB consistently maintains higher performance. Similar trends are observed on MOSEI, where RDIB exhibits a smaller overall performance decline than HME. These results support the effectiveness of RDIB in learning robust multimodal representations under varying noise conditions.

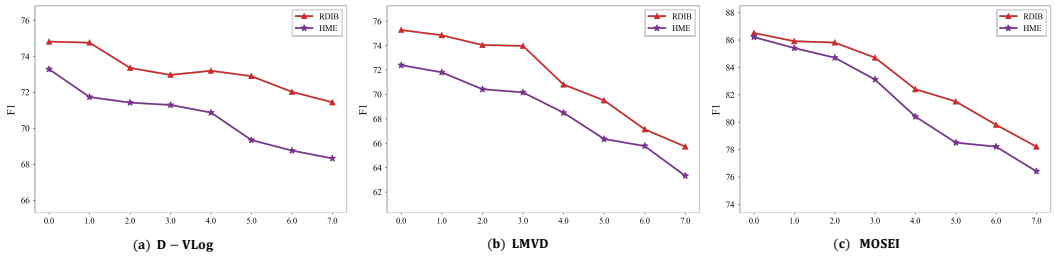


Figure 7: Robustness comparison under different Gaussian noise intensities on D-Vlog, LMVD and MOSEI.

Table 6: Ablation study under fixed and random missing-modality protocols across datasets.

Settings	Full	w/o DIB	w/o L_{target}	w/o L_{task}	w/o Est	w/o DF	w/o RF
Ablations on LMVD							
Vision	**73.10**	71.53	71.94	72.85	69.74	71.90	70.89
Audio	**65.02**	63.00	63.68	63.12	61.36	63.53	62.67
Both	**75.27**	72.08	73.04	73.15	71.94	72.52	72.08
Avg.	**71.13**	68.87	69.55	69.71	67.68	69.32	68.55
Ablations on MOSEI							
MR = 0.1	**84.2/83.7**	82.9/82.3	83.4/82.7	83.2/82.7	83.3/82.8	83.5/82.9	82.6/81.9
MR = 0.4	**77.4/76.8**	76.9/75.1	76.2/74.1	76.9/75.0	76.9/75.2	77.2/75.8	76.4/74.4
MR = 0.7	**72.6/69.1**	70.2/66.8	70.5/67.7	70.2/66.3	70.1/64.5	70.1/65.6	70.5/67.4
Avg.	**78.1/76.5**	76.7/74.7	76.7/74.8	76.8/74.7	76.8/74.2	76.9/74.8	76.5/74.6

D.2 Detailed Ablation Study

To further examine the contribution of each component in RDIB, we report detailed ablation results in Table 6.

On LMVD, removing any component consistently degrades performance across all three fixed missing-modality scenarios. Among the DIB-related ablations, removing the entire DIB causes a larger performance drop than removing either individual constraint, particularly when both modalities are available. This observation supports the complementary contributions of the two constraints. Removing 'Est' produces the largest performance degradation across all three scenarios, highlighting the importance of cross-modal reliability estimation and informative feature reconstruction. Removing 'DF' or 'RF' also consistently reduces performance, demonstrating the contributions of dual-stream and reliability-aware fusion.

On MOSEI, the full RDIB achieves the highest performance across all evaluated missing rates. The effects of individual components vary with the degree of missingness. At lower missing rates, removing 'RF' results in a pronounced performance reduction, while removing the target-modality constraint has a substantial impact at the intermediate missing rate. Under severe missingness (MR = 0.7), removing 'Est' or 'DF' leads to particularly large performance drops, with the second evaluation metric decreasing from 69.1 to 64.5 and 65.6, respectively. These results suggest that reliability estimation and dual-stream fusion become particularly important when available modality information is limited. Overall, the detailed results further support the complementary contributions of the proposed components across different modality-missing conditions.

D.3 Parameter Sensitivity Analysis

We analyze the sensitivity of RDIB to four key hyperparameters on D-Vlog and LMVD, including the loss weights λ_{dib} and λ_{est} in the total optimization objective, and the dual-constraint coefficients β_t and β_y in DIB.

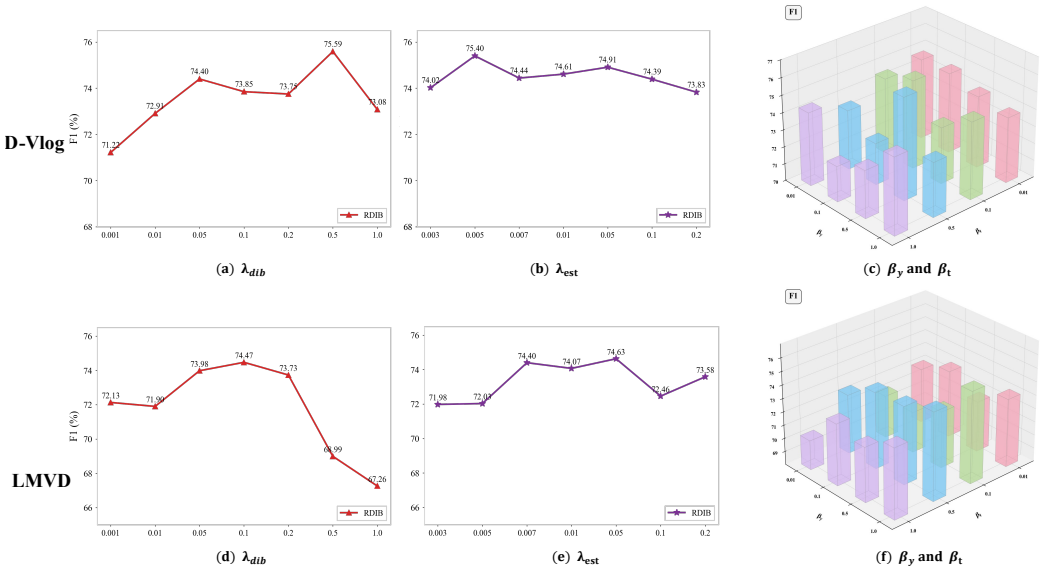


Figure 8: Performance under different hyperparameter values on D-Vlog and LMVD under the fixed missing protocol.

Effect of λ_{dib}. We analyze the influence of the DIB loss weight, as shown in Figure 8. RDIB achieves its highest performance at $\lambda_{dib} = 0.5$ on D-Vlog and $\lambda_{dib} = 0.1$ on LMVD. D-Vlog maintains relatively stable performance across a broad range of settings, whereas LMVD exhibits a more pronounced decline when λ_{dib} becomes excessively large. These results indicate that appropriately balancing DIB optimization with downstream supervision is important, with different datasets exhibiting different sensitivities to information bottleneck regularization.

Effect of λ_{est}. We further evaluate the influence of the reliability estimation loss weight. The highest performance is achieved at $\lambda_{est} = 0.005$ on D-Vlog and $\lambda_{est} = 0.05$ on LMVD. Both datasets maintain competitive performance across several intermediate settings, while excessively small or large values generally degrade performance. This suggests that an appropriate estimation weight is necessary to balance reliability estimation and downstream representation learning.

Effects of dual-constrain weights β_t and β_y. We analyze the joint effects of the two DIB constraint coefficients, as shown in Figure 8. RDIB achieves the highest F1-score on D-Vlog at $\beta_t = 0.5$ and $\beta_y = 0.5$, while LMVD reaches its peak performance at $\beta_t = 0.1$ and $\beta_y = 1.0$. These results indicate that the two constraints require dataset-specific balancing, with LMVD benefiting from relatively stronger task-relevant supervision. Moreover, excessively increasing either coefficient does not consistently improve performance, highlighting the importance of jointly balancing target-modality information preservation and task relevance.

D.4 Model Complexity Analysis

We compare the computational complexity of RDIB with representative state-of-the-art methods in terms of total parameters, FLOPs, and inference time, as reported in Table 7. For LMVD, we evaluate the longest sample to examine computational efficiency under long-sequence conditions. RDIB achieves the lowest parameter count, with substantially fewer FLOPs than DepMamba. Although P^{3}HF requires fewer FLOPs, RDIB achieves considerably faster inference, demonstrating a favorable balance between computational cost and inference efficiency for long-sequence inputs.

For MOSEI, we randomly select a sample for evaluation with each model configured using the hyperparameter settings reported in its original paper. RDIB requires fewer parameters than HME, while maintaining a parameter count close to IMDer. Despite having slightly higher FLOPs than both baselines, its inference time remains between those of IMDer and HME. Overall, these results demonstrate that RDIB maintains competitive computational efficiency across different multimodal learning scenarios.

Table 7: Computational complexity comparison on LMVD and MOSEI in terms of model parameters, FLOPs, and inference time.

Models	Total Parameters	Total FLOPs	Inference Time
Comparison on LMVD			
DepMamba	2.43M	14.76G	15.90 ms
P^3HF	9.81M	0.84G	149.64 ms
Ours	1.52M	3.41G	45.93 ms
Comparison on MOSEI			
IMDer	122.95M	8.75G	49.68 ms
HME	145.58M	9.07G	37.76 ms
Ours	124.95M	9.85G	43.84 ms

D.5 Further Analysis

The effectiveness of RDIB largely depends on the variational approximation adopted in DIB, where the latent posterior is parameterized as a diagonal Gaussian distribution. Although this assumption enables tractable optimization and uncertainty estimation, its limited expressiveness may introduce approximation errors when the underlying latent distributions exhibit long-tailed characteristics. Since modality reliability is derived from posterior uncertainty, such distributional mismatch may further affect reliability estimation and subsequent multimodal fusion. Future work will investigate more expressive variational posterior families to improve distributional adaptability and uncertainty estimation under diverse multimodal conditions.