DYNAMIC MODALITY-AWARE DUAL-NOISE TRAINING FOR ROBUST MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Sentiment Analysis (MSA) aims to infer human emotions by integrating text, audio, and visual modalities. However, real-world deployment remains challenging due to heterogeneous perturbations, such as random data loss and transmission-induced feature distortion in distributed sensing environments. Existing methods frequently presuppose static modality dominance and focus on mitigating isolated forms of noise. To address the diverse and complex noise perturbations encountered in real-world scenarios, we propose a Dynamic Modality-Aware Dual-Noise Training framework (DMDN). Our framework first introduces the Dynamic Information-guided Modality Correction Module (DIMC), which estimates modality degradation via KL divergence and dynamically identifies the most reliable modality to enhance corrupted features. In this framework, modality features are encoded into sparse spike trains by a spiking neuron-based sender, transmitted through simulated noisy channels, and decoded by a State Space Model (SSM)-based receiver. Our framework jointly models random data missing and channel noise, better reflecting the complex perturbation patterns in real-world distributed systems. Empirically, joint training under such multi-noise conditions encourages the model to suppress irrelevant perturbations while preserving task-relevant information. Extensive experiments demonstrate that our approach consistently outperforms state-of-the-art baselines across multiple MSA benchmarks.

1 INTRODUCTION

Multimodal Sentiment Analysis (MSA) has emerged as a crucial task in affective computing, aiming to infer human emotions and sentiments by integrating heterogeneous information sources such as text, audio, and visual modalities. MSA holds significant potential for a wide array of applications, including human-computer interaction, social media analytics, and mental health assessment. To handle incomplete or noisy modalities, several approaches have been proposed: Li et al. (2024a) propose a unified missing-modality distillation framework, Yuan et al. (2023) train adversarially against noise-perturbed samples, and Zhang et al. (2024) introduce random data missing scenarios with a comprehensive evaluation on three widely used datasets.

Meanwhile, as deep learning models grow in scale, collaborative intelligence (CI) (Li et al., 2023a; 2024b) has emerged as a promising paradigm that partitions computation between edge devices and the cloud. In such distributed settings, perturbations arise from multiple interacting mechanisms. For example, modality features may be lost at the acquisition stage due to sensor failures, while unstable wireless channels introduce transmission noise that causes irreversible, unstructured feature distortions. Conventional strategies that address only random data missing are inadequate here. Our work targets a more realistic MSA setting where models must simultaneously contend with random data loss and feature corruption, and existing approaches face two significant challenges:

1. Modality Dominance Bias: Most completion strategies in MSA follow a static language-dominant paradigm (Zhang et al., 2024). However, under random data missing, textual features are often highly informative but less redundant, so losing a single emotionally salient keyword can drastically shift the predicted sentiment. Moreover, each modality's contribution is inherently dynamic.

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.

For instance, during intense emotion, prosodic cues in audio or subtle facial micro-expressions may convey stronger emotional signals than text.

2. Heterogeneous Perturbation Granularity: Most robustness-focused MSA methods address a single noise type, typically feature-level random dropout. Real-world perturbations, however, are multi-type and multi-scale: random data missing disrupts the macro-structural composition, while channel noise introduces micro-level semantic distortions during transmission. This motivates an anti-noise training framework that optimizes representation learning under multi-scale disturbances.

To address these challenges, we propose DMDN. First, the Dynamic Information-guided Modality Correction Module (DIMC) estimates each modality's degradation during training via the KL divergence between its original and randomly masked features, dynamically selects the least degraded modality as dominant, and uses the remaining modalities to correct and enhance it; during evaluation, a jointly trained feature reconstructor restores the incomplete data. Second, to address channel noise in distributed multimodal systems, we separately model random data missing and channel-level perturbations, and propose a joint source-channel coding framework for anti-noise training. In this framework, continuous modality features are encoded into sparse binary spike trains by a sender based on a spiking neural network, transmitted through a simulated noisy channel, and decoded by a receiver based on a state space model. The spike-based sender enables efficient and robust transmission, while the state space model recovers channel-corrupted features through temporal modeling. The three main contributions can be summarized as:

- We propose a Dynamic Information-guided Modality Correction (DIMC) module, which dynamically enhances cross-modal feature completion under diverse missing-data patterns.
- We develop a joint source-channel coding anti-noise training framework, where modality features are encoded into spike trains via a spiking-neuron-based sender and decoded by a State Space Model receiver, enabling reliable transmission in noisy distributed systems.
- Extensive experiments across multiple MSA benchmarks demonstrate that our framework consistently outperforms existing methods and achieves state-of-the-art performance.

2 Related Work

2.1 Robust Multimodal Sentiment Analysis

Robust multimodal sentiment analysis (MSA) aims to maintain reliable sentiment prediction when multimodal inputs are noisy, incomplete, or corrupted (Yang et al., 2022; Liang et al., 2020; Rahman et al., 2020). To improve robustness, existing studies mainly focus on modeling intra- and inter-modal interactions, learning complementary cross-modal representations, estimating modality reliability, and suppressing unreliable modality information. Representative approaches include adversarial robustness learning (Yuan et al., 2023), language-dominated correction under random missing conditions (Zhang et al., 2024), shared-private representation learning (Hazarika et al., 2020), self-supervised multimodal consistency learning (Yu et al., 2021), multiplicative fusion with modality effectiveness estimation (Mittal et al., 2020), and transformer-based robust fusion methods (Tsai et al., 2019; Zhang et al., 2023). These methods have substantially improved robust MSA performance under incomplete or noisy inputs, but most are developed for a single disturbance type, especially random data missing, and rarely consider feature corruption caused by transmission noise.

Besides these robustness-oriented MSA methods, spiking neural networks (SNNs) have recently shown growing potential in multimodal learning due to their event-driven and temporally sparse computation (Yao et al., 2023; Fang et al., 2023). Their spike-based representation is naturally suited for temporally structured multimodal signals, and prior studies have demonstrated promising results in multimodal understanding and recognition tasks (Li et al., 2023b;c). This suggests that SNNs can provide an effective and efficient way to model multimodal dynamics. However, existing SNN-based multimodal methods mainly emphasize representation learning or recognition performance, rather than robustness in sentiment analysis under heterogeneous perturbations.

In contrast, our work considers multimodal inputs that are simultaneously affected by source-side missing data and channel-side transmission noise, and combines dynamic modality correction with spike-based source-channel transmission. By jointly modeling disturbances at both the input and transmission levels, the proposed framework enables robust MSA under multi-scale degradation.

3 METHOD

3.1 PROBLEM FORMULATION ON DUAL-NOISE MODELING

We consider multimodal sentiment analysis with three modalities, namely language, audio, and vision, denoted by $\mathcal{X} = \{X_l, X_a, X_v\}$. In real-world scenarios, multimodal inputs are often affected by two types of noise: *random data missing* and *source-channel transmission noise*. The former reduces semantic completeness at the input level, while the latter corrupts individual bits.

Random Data Missing. We model source-side missing corruption as a pre-processing operation applied before the inputs are fed into the model. For each modality $m \in \{l, a, v\}$, let $X_m = [x_{m,1}, x_{m,2}, \ldots, x_{m,L_m}]$ denote the input sequence, where L_m is the number of tokens or temporal units. We introduce a binary keep mask $z_m \in \{0,1\}^{L_m}$, where $z_{m,t} = 1$ means that the t-th unit is retained and $z_{m,t} = 0$ means that it is missing.

For each modality $m \in \{l, a, v\}$, missing corruption is applied by replacing the missing unit with a modality-specific default token:

$$\tilde{x}_{m,t} = z_{m,t} x_{m,t} + (1 - z_{m,t}) r_m, \tag{1}$$

where r_m denotes the replacement for missing inputs. For acoustic and visual modalities, $r_m = \mathbf{0}$ is a zero vector with the same dimensionality as $x_{m,t}$; for the language modality, $r_l = e_{\texttt{[UNK]}}$ is the embedding of the special token `[UNK]`. The corrupted sequence can thus be written uniformly as

$$\tilde{X}_m = Z_m \odot X_m + (1 - Z_m) \odot R_m, \tag{2}$$

where Z_m is the keep mask broadcast along the feature dimension, R_m is the broadcast form of r_m, and $\odot$ denotes element-wise multiplication.

The mask variables are independently sampled from Bernoulli distributions:

$$z_{m,t} \sim \text{Bernoulli}(\rho_m), \quad m \in \{l, a, v\}, \tag{3}$$

where $\rho_m \in [0,1]$ is the keep probability of modality m. This formulation provides a unified description of source-side incompleteness across different modalities.

Source-Channel Transmission Noise. After source-side corruption (and DIMC correction, Sect. 3.2), features are encoded into binary spike trains by the SNN sender (Sect. 3.3) and transmitted through unreliable digital channels. We consider two standard memoryless channels: the Binary Symmetric Channel (BSC) and Binary Erasure Channel (BEC).

Let $b \in \{0,1\}$ denote a transmitted bit and $\hat{b}$ denote its received version.

Binary Symmetric Channel (BSC). Under the BSC model, each transmitted bit is independently flipped with probability $\epsilon \in [0,1]$:

$$\hat{b} = b \oplus n, \quad n \sim \text{Bernoulli}(\epsilon), \tag{4}$$

where $\oplus$ denotes the XOR operation. Equivalently,

$$P(\hat{b} = b) = 1 - \epsilon, \qquad P(\hat{b} \neq b) = \epsilon. \tag{5}$$

Binary Erasure Channel (BEC). Under the BEC model, each transmitted bit is independently erased with probability $\delta \in [0,1]$:

$$\hat{b} = (1 - z)b + ze, \quad z \sim \text{Bernoulli}(\delta), \tag{6}$$

where e denotes an erasure symbol, indicating that the receiver knows the bit is missing but cannot determine whether it was 0 or 1. Equivalently,

$$P(\hat{b} = b) = 1 - \delta, \qquad P(\hat{b} = e) = \delta. \tag{7}$$

In our implementation, the erasure symbol e is represented by the constant 0.5.

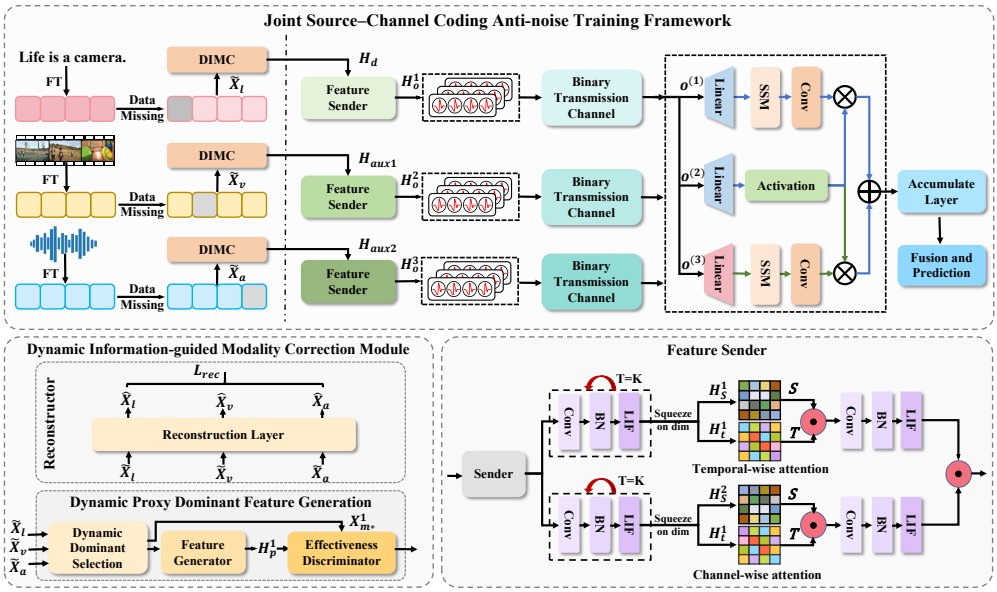

Figure 1: Overview of the proposed Dynamic Modality-Aware Dual-Noise Training framework. The inputs from textual, visual, and acoustic modalities first undergo random data missing, followed by repair and enhancement by DIMC. Each modality-specific feature is then input into an SNN-based feature sender, encoded into sparse spike trains and transmitted through simulated binary channels. The SSM reconstructs the noise-corrupted features, which are fused across modalities via an accumulation layer for final sentiment prediction.

3.2 Dynamic Information-guided Modality Correction Module

Existing methods for incomplete multimodal analysis often assume that the language modality is consistently dominant. Under random corruption, however, the most reliable modality varies across samples, since different modalities suffer different degrees of information loss. We estimate the degradation severity of each modality and adaptively select the least corrupted modality as dominant.

During training, both the clean modality feature $X_m \in \mathbb{R}^{B\times L\times D}$ and its degraded counterpart $\tilde{X}_m$ are available. For each sample index n and temporal position t, we treat the corresponding D-dimensional feature vector as a discrete distribution over the feature dimension. Specifically, we normalize the feature vector by applying a softmax over the last dimension:

$$P(x_m^{(n,t)})_i = \frac{\sigma(x_m^{(n,t)})_i + \varepsilon}{\sum_{j=1}^{D}\left(\sigma(x_m^{(n,t)})_j + \varepsilon\right)}, \tag{8}$$

where ε is a small positive constant for numerical stability, $\sigma(\cdot)$ denotes the softmax operation, and i, j index the feature dimension.

We then compute the token-wise degradation score at each token position using the KL divergence between the clean and degraded distributions:

$$s_m^{(n,t)} = \sum_{i=1}^{D} P(x_m^{(n,t)})_i \log \frac{P(x_m^{(n,t)})_i}{P(\tilde{x}_m^{(n,t)})_i}. \tag{9}$$

The degradation score for modality m in sample n is obtained by averaging the token-wise KL scores over the temporal dimension:

$$s_m^{(n)} = \frac{1}{L}\sum_{t=1}^{L} s_m^{(n,t)}. \tag{10}$$

A larger value of $s_m^{(n)}$ indicates more severe feature degradation; we use it as a proxy for modality reliability. During inference, the clean modality features are unavailable. We therefore introduce

a modality-shared reconstructor R, implemented as a lightweight single Transformer layer, which produces a pseudo-clean feature $\hat{X}_m = R(\tilde{X}_m)$. The degradation severity is then estimated via KL divergence between $\hat{X}_m$ and $\tilde{X}_m$, and the modality with the lowest degradation becomes dominant.

The remaining two modalities are then used to generate a proxy dominant representation $\hat{H}_p$, and a completeness score w estimated from the degraded dominant feature balances this proxy with the original degraded feature H_m^*:

$$H_d = (1 - \hat{w}) \cdot \hat{H}_p + \hat{w} \cdot H_m^*, \tag{11}$$

3.3 Joint Source-Channel Coding Anti-noise Training Framework

To enhance the encoding of multimodal information and to improve robustness against transmission perturbations, we propose a JSCC Anti-Noise Training Framework. This framework consists of a feature sender, a noisy binary transmission channel (Sect. 3.1), and a feature receiver.

3.3.1 Feature Sender

The feature sender encodes the modality feature $\mathbf{H}_m \in \mathbb{R}^{B\times L\times D}$ into a binary spike train $\mathbf{H}_{\text{enc}} \in \{0,1\}^{T_{\text{snn}}\times B\times L\times D}$. Here, T_{snn} denotes the number of SNN simulation timesteps, distinct from the token sequence length L. We design a three-branch SNN-based sender, where each branch is a Conv-BN-LIF block with a different kernel size, yielding multi-scale temporal features and discrete spike sequences for subsequent transmission:

$$\mathbf{H}_{\text{enc}} = \text{LIF}(\text{BN}(\text{Conv}_k(\mathbf{H}_m))), \quad k \in \{5, 3, 1\} \tag{12}$$

where $\text{Conv}_k(\cdot)$ represents the convolution operation with the kernel size of k and $\text{BN}(\cdot)$ is the batch normalization operation. The dynamics equation of the LIF neuron can be written as:

$$\tau_m \frac{dV(t)}{dt} = -(V(t) - V_{rest}) + RI(t), \tag{13}$$

where $V(t)$ denotes the membrane potential, V_{rest} is the resting potential, R is the membrane resistance, $I(t)$ represents the input current, and τ_m is the membrane time constant. A spike is emitted when $V(t)$ exceeds the threshold V_{th}, after which the membrane potential is reset.

3.3.2 Temporal Sequence-Aware Attention Module.

To better model dependencies in spiking sequences, we introduce a Temporal Sequence-Aware Attention (TSA) module (Zhu et al., 2024) after the Conv-BN-LIF layer. Given the spike representation $\mathbf{H}_{enc} \in \{0,1\}^{T_{\text{snn}}\times B\times L\times D}$, we first apply global average pooling over the feature dimension to obtain a compact representation $\mathbf{H}_q \in \mathbb{R}^{T_{\text{snn}}\times B\times L}$. We then reshape it as $\mathbf{H}' \in \mathbb{R}^{B\times T_{\text{snn}}\times L}$, which is fed to both the temporal and sequence attention branches.

Here, t indexes SNN simulation timesteps and l indexes token positions. In the temporal branch, attention is generated by a 1D convolution along the temporal dimension:

$$T_{b,t,l} = \sum_{i=0}^{k_t-1} W_{i,l}^{(T)}\, H'_{b,\ t+i-\lfloor k_t/2 \rfloor,\ l}, \tag{14}$$

where $W^{(T)}$ is a learnable temporal convolution kernel with kernel size k_t.

In the sequence branch, attention is generated by a 1D convolution along the sequence dimension:

$$S_{b,t,l} = \sum_{j=0}^{k_l-1} W_{j,t}^{(L)}\, H'_{b,\ t,\ l+j-\lfloor k_l/2 \rfloor}, \tag{15}$$

where $W^{(L)}$ is a learnable sequence convolution kernel of size k_l, with zero-padding at both ends of the sequence.

The two attention maps have shape $\mathbb{R}^{B\times T_{\text{snn}}\times L}$. We fuse them by element-wise multiplication and normalization, then explicitly permute the resulting map to $\mathbf{A}^{\pi} \in \mathbb{R}^{T_{\text{snn}}\times B\times L}$ before broadcasting

it along the feature dimension. This aligned map is applied to the original spike tensor through residual gating:

$$\mathbf{H}_o = w_r \cdot \left(\mathbf{H}_{enc} \odot \mathbf{A}^{\pi}\right) + (1 - w_r) \cdot \mathbf{H}_{enc}, \tag{16}$$

where $\mathbf{A} = \psi(\mathbf{T} \odot \mathbf{S})$ and $\mathbf{A}^{\pi}_{t,b,l} = \mathbf{A}_{b,t,l}$, w_r is a learnable scalar, and $\psi(\cdot)$ denotes a normalization function. In this way, TSA emphasizes informative temporal and sequence-level patterns while preserving the original spiking representation through the residual path. The refined features are then further processed by another Conv-BN-LIF unit before transmission.

3.3.3 Feature Receiver

Upon reception, the noisy spiking sequence is first processed by a Selective State Space Model (SSM) (Gu & Dao, 2023), which exploits temporal dependencies to suppress channel noise and restore feature consistency. The denoised features are then refined by a lightweight gating-based fusion module for feature reconstruction:

$$\mathbf{c}_t = \phi\Big(\mathbf{o}_t^{(5)} \odot \mathbf{o}_t^{(3)}\Big)\, \mathbf{o}_t^{(1)} + \mathbf{o}_t^{(1)}, \tag{17}$$

where $\mathbf{o}_t^{(5)}$, $\mathbf{o}_t^{(3)}$, and $\mathbf{o}_t^{(1)}$ denote the features generated by the 5×5, 3×3, and 1×1 convolution branches, and $\phi(\cdot)$ denotes the softmax normalization function.

The refined spike sequence $\mathbf{c} \in \mathbb{R}^{T_{\text{snn}} \times B \times L \times D}$ is then averaged over the SNN simulation timesteps to obtain the firing-rate representation:

$$\mathbf{F} = \frac{1}{T_{\text{snn}}} \sum_{t=1}^{T_{\text{snn}}} \mathbf{c}_t, \tag{18}$$

where $\mathbf{F} \in \mathbb{R}^{B \times L \times D}$ is a real-valued feature tensor used by the downstream modules.

Multimodal Fusion and Prediction. After feature reception, we adopt the Adaptive Hypermodality Learning strategy to hierarchically fuse information across modalities. A 4-layer Transformer encoder and a classification head are then used for final multimodal prediction. Additional architectural details follow (Zhang et al., 2023).

3.4 Training Objectives

Our training objective consists of four components: a prediction loss for sentiment regression, a reconstruction loss for the modality reconstructor, an adversarial loss for the DIMC module, and a weighting loss for learning the fusion coefficient. These losses are defined as

$$\begin{aligned} &\mathcal{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{y}^{(i)} - y^{(i)}\|_2^2, \quad \mathcal{L}_{\text{rec}} = \frac{1}{3N}\sum_{i=1}^{N}\sum_{m\in\{l,a,v\}} \|\hat{X}_m^{(i)} - X_m^{(i)}\|_2^2, \\ &\min_{\theta_D}\max_{\theta_{\text{DFG}}} \mathcal{L}_{\text{adv}} = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_p^{(i)\top} \log D(\hat{H}_p^{(i)}; \theta_D), \quad \mathcal{L}_{\text{w}} = \frac{1}{N}\sum_{i=1}^{N} \|w^{(i)} - \hat{w}^{(i)}\|_2^2, \end{aligned} \tag{19}$$

where N is the number of training samples, $\hat{y}$ and y denote the predicted and ground-truth sentiment labels, $\hat{X}_m$ and X_m denote the reconstructed and complete features of modality m, $\hat{H}_p$ denotes the generated dominant-modality feature, $\mathbf{y}_p$ denotes the source label for the adversarial classification, and $\hat{w}$ and w denote the predicted completeness weight and its supervision target, respectively. Here, $D(\cdot)$ denotes the effectiveness discriminator, θ_D denotes its parameters, and θ_{DFG} denotes the parameters of the proxy dominant feature generator. The discriminator D is implemented as a lightweight two-layer MLP that predicts whether a fused representation is produced by the generator or taken from the genuine dominant modality. It is optimized alternately with the generator under the min-max objective in Eq. (19), which encourages the generated proxy features to be indistinguishable from real dominant-modality representations. The overall objective is

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{pred}} + \lambda_2\mathcal{L}_{\text{rec}} + \lambda_3\mathcal{L}_{\text{adv}} + \lambda_4\mathcal{L}_{\text{w}}. \tag{20}$$

Table 1: Results comparison on MOSI and MOSEI datasets. $\text{DMDN}_{\text{STD}/\text{BSC}/\text{BEC}}$ represents the model training without channel noise, with channel noise on BSC and BEC channels, respectively. For the slash-separated Acc-2 and F1 values, the first value is the Non0 setting, which excludes neutral samples ($y = 0$) and classifies positive ($y > 0$) versus negative ($y < 0$); the second is the Has0 setting, which classifies all samples as non-negative ($y \geq 0$) versus negative ($y < 0$).

MOSI						
Method	Acc-7 ↑	Acc-5 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
self-MM (Yu et al., 2021)	26.97	31.39	67.43/67.54	64.27/66.81	1.129	0.503
MMIM (Han et al., 2021)	28.23	29.89	68.09/66.52	66.15/64.59	1.128	0.501
CENET (Wang et al., 2022)	28.33	30.90	72.46/66.08	71.10/63.50	1.130	0.496
TETFN (Wang et al., 2023)	31.92	35.71	72.10/70.85	72.16/70.81	1.091	0.516
ALMT (Zhang et al., 2023)	28.42	31.25	68.24/65.94	69.74/68.54	1.138	0.485
LNLN (Zhang et al., 2024)	33.53	37.03	73.56/72.99	73.63/72.91	1.096	0.518
MFMB-Net (Tao et al., 2025)	28.57	30.65	66.22/66.25	66.30/66.21	1.205	0.453
MRUF (Ma et al., 2026)	31.05	35.57	69.97/70.08	69.83/70.08	1.138	0.516
DMDN_{STD}	33.09	37.46	**74.24**/73.18	74.44/73.44	1.081	**0.519**
DMDN_{BSC}	34.40	38.48	74.09/72.45	**74.49**/72.59	**1.067**	**0.519**
DMDN_{BEC}	**35.57**	**39.50**	72.16/**73.63**	72.19/**73.79**	1.084	0.518
MOSEI						
Method	Acc-7 ↑	Acc-5 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
self-MM (Yu et al., 2021)	42.70	43.14	71.97/75.81	67.40/70.38	0.733	0.477
MMIM (Han et al., 2021)	38.68	39.21	71.75/74.45	67.70/67.96	0.775	0.470
CENET (Wang et al., 2022)	45.12	45.52	73.33/77.16	69.80/74.14	0.720	0.515
TETFN (Wang et al., 2023)	46.17	47.05	76.97/72.89	77.15/73.96	0.686	0.565
ALMT (Zhang et al., 2023)	37.82	38.34	77.40/77.48	77.73/77.80	0.683	0.461
LNLN (Zhang et al., 2024)	45.83	45.41	78.94/**78.58**	79.09/79.12	0.674	0.556
MFMB-Net (Tao et al., 2025)	42.98	43.73	47.64/58.11	49.44/60.02	0.754	0.568
MRUF (Ma et al., 2026)	46.49	47.28	77.68/77.38	76.05/77.38	0.975	0.562
DMDN_{STD}	**50.83**	**50.51**	**78.98**/78.23	79.99/78.17	**0.633**	**0.582**
DMDN_{BSC}	46.00	47.18	77.24/77.91	77.17/**79.26**	0.684	0.577
DMDN_{BEC}	45.12	45.89	77.44/78.30	**80.13**/77.53	0.688	0.578

Table 2: Results comparison on SIMS dataset. $\text{DMDN}_{\text{STD}/\text{BSC}/\text{BEC}}$ represents the model training without channel noise, with channel noise on BSC and BEC channels, respectively.

Method	SIMS					
	Acc-5 ↑	Acc-3 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
self-MM (Yu et al., 2021)	32.02	53.90	71.41	67.11	0.517	0.390
MMIM (Han et al., 2021)	33.41	52.37	68.49	64.81	0.553	0.336
CENET (Wang et al., 2022)	23.12	54.05	68.71	57.92	0.588	0.107
TETFN (Wang et al., 2023)	31.29	53.61	64.33	65.62	0.538	0.368
ALMT (Zhang et al., 2023)	18.38	47.12	68.27	71.22	0.563	0.395
LNLN (Zhang et al., 2024)	36.32	59.08	74.40	77.59	0.507	0.415
MFMB-Net (Tao et al., 2025)	26.79	64.96	78.79	79.45	0.464	0.584
MRUF (Ma et al., 2026)	33.26	61.71	73.52	73.83	0.491	0.494
DMDN_{STD}	40.70	**67.61**	**79.43**	**80.99**	**0.447**	**0.597**
DMDN_{BSC}	40.92	66.96	79.21	79.66	0.453	0.589
DMDN_{BEC}	**42.01**	67.17	78.12	78.04	0.456	**0.597**

4 EXPERIMENTS

4.1 DATASETS

We conduct experiments on three widely used multimodal sentiment analysis benchmarks: MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and SIMS (Yu et al., 2020).MOSI contains 2,199 sentiment-labeled utterances from YouTube videos. MOSEI is a larger benchmark with over 23,000 annotated utterances from more than 1,000 speakers. SIMS is a Chinese multimodal sentiment dataset collected from interactive scenarios such as dialogues and interviews.

4.2 IMPLEMENTATION DETAILS

The experiments are conducted on a platform equipped with an Intel(R) Xeon(R) Platinum 8481 CPU and an NVIDIA 4090D GPU. The projected token sequence length is set to $L = 8$, the number of SNN simulation timesteps to $T_{\text{snn}} = 8$, and the feature dimension to $D = 128$. The model is trained with a batch size of 64 and an initial learning rate of 1×10^{-4}, using the AdamW optimizer. For the multi-task loss weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, we set them to $0.9/0.8/0.1/1.0$ for both the MOSI and MOSEI datasets. For the SIMS dataset, λ_2 is adjusted to 0.6 to accommodate dataset-specific characteristics. All models are trained for 200 epochs.

Table 3: Effects of different components on MOSI dataset.

Method	Acc-7 ↑	Acc-5 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
DMDN_{BSC}	**34.40**	**38.48**	**72.45**	**72.59**	**1.0670**	0.5190
w/o DIMC	33.82	37.32	72.30	72.54	1.1045	0.4844
w/o TSA	32.80	36.15	70.33	70.41	1.1203	0.4943
w/o SSM	34.15	38.34	70.85	70.80	1.0754	**0.5314**
w/o DIMC&TSA&SSM	31.92	35.71	69.83	70.11	1.1220	0.4780
SNN-to-Conv	31.13	34.32	71.25	71.15	1.1485	0.4658
SSM-to-Trans	32.15	35.28	71.82	71.13	1.1386	0.4781

4.3 RESULTS COMPARISON

Tables 1 and 2 report the performance of our model on **MOSI**, **MOSEI**, and **SIMS** under a missing rate of 0.5, with BSC bit-flip probability epsilon=0.2 for channel-noise settings. We evaluate three variants: DMDN_{STD} without channel noise and DMDN_{BSC}/DMDN_{BEC} with BSC/BEC noise, respectively. Results under joint BSC+BEC noise are provided in the appendix.

On MOSI, DMDN_{STD} consistently outperforms the strongest baseline LNLN on the main classification metrics, improving Acc-2/F1 from 73.56/73.63 to 74.24/74.44 and Acc-5 from 37.03 to 37.46, while also achieving lower MAE (1.081 vs. 1.096) and higher correlation (0.519 vs. 0.518). On MOSEI, DMDN_{STD} achieves the best Acc-7, Acc-5, MAE, and correlation, reaching 50.83, 50.51, 0.633, and 0.582, respectively, showing a favorable trade-off across heterogeneous evaluation metrics. When additional channel noise is introduced during training, the model remains highly competitive and, in several cases, performs even better than DMDN_{STD}. On MOSI, DMDN_{BSC} and DMDN_{BEC} further improve Acc-7 and Acc-5 over the standard setting, indicating that joint training with random missing and channel perturbation can enhance robustness. A similar trend is observed on MOSEI, where DMDN_{BSC} achieves the best Has0 F1 (79.26) and DMDN_{BEC} attains the best Non0 F1 (80.13), while DMDN_{STD} and LNLN obtain the best Non0 and Has0 Acc-2 (78.98 and 78.58), respectively. This suggests that bit-level transmission perturbation may act as a form of structured regularization on MOSI, encouraging more noise-tolerant representations. On SIMS, all three DMDN variants achieve strong and stable performance, consistently outperforming prior methods on most metrics. DMDN_{STD} achieves the best Acc-3, Acc-2, F1, and MAE, with scores of 67.61, 79.43, 80.99, and 0.447, respectively, while achieving a tied-best correlation of 0.597. DMDN_{BEC} further achieves the best Acc-5 of 42.01, and DMDN_{BSC} remains highly competitive across all metrics. These results show that DMDN maintains strong robustness on SIMS under both missing-modality and channel-noise conditions.

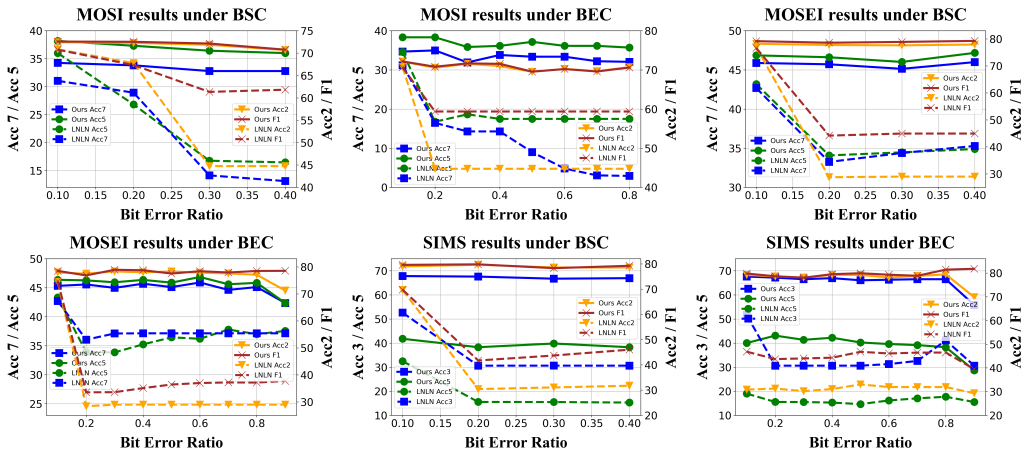

Figure 2: Robustness evaluation under varying BSC bit-flip probabilities and BEC erasure probabilities on three multimodal sentiment datasets.

4.4 Robustness Analysis

As shown in Fig. 2, DMDN consistently achieves better performance than LNLN under both BSC and BEC noise on MOSI, MOSEI, and SIMS, with smoother degradation as the error rate increases, whereas LNLN tends to exhibit lower performance and less stable behavior in several settings, indicating weaker robustness to channel corruption. DMDN also shows stronger resistance to the cliff effect: its performance remains relatively stable under BSC noise, and under BEC noise only MOSI exhibits a moderate drop around an erasure probability of 0.3, while MOSEI and SIMS stay comparatively stable even at high error rates.

4.5 Ablation Study

Table 3 summarizes ablation results on MOSI under BSC noise (BER $= 0.2$). Removing DIMC reduces Acc-7 and Acc-5 from 34.40/38.48 to 33.82/37.32, while also lowering F1 and correlation, confirming the benefit of dynamic modality correction. Removing TSA leads to a clearer degradation in Acc-2 and F1, showing that temporal modeling in the sender is important for robust prediction. Removing SSM also decreases Acc-2 and F1, though it slightly raises correlation, suggesting the receiver mainly improves classification consistency rather than uniformly benefiting every metric. Removing all three modules causes the largest overall performance drop, indicating that these components provide complementary gains.

To further justify the architectural design, we replace the SNN sender with a non-spiking convolutional alternative (SNN-to-Conv) and replace the SSM receiver with a Transformer decoder (SSM-to-Trans). Both substitutions underperform the full model across most metrics, especially Acc-7, Acc-5, F1, and MAE, and also lead to clear drops in correlation. These results support the effectiveness of spike-based encoding and SSM-based denoising. An additional ablation on dominance selection is provided in the appendix (Table 7).

5 Conclusion

In this work, we presented DMDN for robust multimodal sentiment analysis. By explicitly modeling both random data missing and channel transmission noise, our approach departs from static modality assumptions and single-noise robustness paradigms. Through the introduction of the DIMC module and a biologically inspired spiking-based source-channel coding architecture, our work enables sample-wise modality correction and resilient signal reconstruction. Extensive experiments demonstrate that DMDN outperforms baselines under complex perturbation scenarios, highlighting the potential of noise-aware frameworks for reliable multimodal AI in real-world deployment.

AI USE STATEMENT

In the process of developing this work, we utilized LLMs for grammar correction and occasionally as a rewriting tool. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Wei Fang, Yanqi Chen, Jianhao Ding, Zhaofei Yu, Timothee Masquelier, Ding Chen, Liwei Huang, Huihui Zhou, Guoqi Li, and Yonghong Tian. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40):eadi1480, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Wei Han, Hui Chen, and Soujanya Poria. 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*, pp. 9180–9192, 2021.

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.

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

Wenrui Li, Zhengyu Ma, Liang-Jian Deng, Xiaopeng Fan, and Yonghong Tian. Neuron-based spiking transmission and reasoning network for robust image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3516–3528, 2023a. doi: 10.1109/TCSVT.2022.3233042.

Wenrui Li, Zhengyu Ma, Liang-Jian Deng, Hengyu Man, and Xiaopeng Fan. Modality-fusion spiking transformer network for audio-visual zero-shot learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 426–431. IEEE, 2023b.

Wenrui Li, Xi-Le Zhao, Zhengyu Ma, Xingtao Wang, Xiaopeng Fan, and Yonghong Tian. Motion-decoupled spiking transformer for audio-visual zero-shot learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3994–4002, 2023c.

Wenrui Li, Jiahui Li, Mengyao Ma, Xiaopeng Hong, and Xiaopeng Fan. Multi-scale spiking pyramid wireless communication framework for food recognition. *IEEE Transactions on Multimedia*, pp. 1–13, 2024b. doi: 10.1109/TMM.2024.3368964.

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

Haoran Ma, Yinfeng Yu, and Liejun Wang. Mruf: Multi-granularity routing with uncertainty-aware fusion for robust multimodal sentiment analysis. *IEEE International Conference on Systems, Man, and Cybernetics*, 2026.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1359–1367, 2020.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2020, pp. 2359, 2020.

Chuanqi Tao, Jiaming Li, Tianzi Zang, and Peng Gao. A multi-focus-driven multi-branch network for robust multimodal sentiment analysis. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, 2025.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, pp. 6558, 2019.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, pp. 4909–4921, 2022.

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

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

Man Yao, JiaKui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64043–64058. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ca0f5358dbadda74b3049711887e9ead-Paper-Conference.pdf.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 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*, pp. 3718–3727, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10790–10797, 2021.

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

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

Amir 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*, pp. 2236–2246, 2018.

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

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

Rui-Jie Zhu, Malu Zhang, Qihang Zhao, Haoyu Deng, Yule Duan, and Liang-Jian Deng. TCJA-SNN: Temporal-channel joint attention for spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

A APPENDIX

A.1 EXTENDED STUDY UNDER MIXED NOISE

To further assess the robustness of the proposed framework in more challenging and realistic communication scenarios, we extend our experiments to include mixed noise conditions, where both BSC and BEC corruptions are applied. This setting simulates compound degradations during feature transmission, in which features may simultaneously suffer from bit flips and erasures, providing a more stringent test of model resilience.

For the mixed-channel experiments, the single noise level p is used for both channels: the BEC erasure probability is set to $\delta = p$ and the BSC bit-flip probability is set to $\epsilon = p$. The two channels are applied sequentially, with BEC applied first and BSC applied afterward.

Quantization and channel interface. The two model paths use a matched 3-bit channel interface. In DMDN, the projected feature $H_m \in \mathbb{R}^{B\times L\times D}$ is repeated over $T_{\text{snn}} = 8$ simulation steps and processed by the three Conv–BN–LIF sender branches to produce the spike representation. Before entering the discrete channel, the channel symbols are quantized to integers and serialized with `num_bits=3`; BSC/BEC masks are then applied to the serialized bits. The received 3-bit groups are decoded and passed to the SSM receiver, which averages the recovered representation over simulation steps. The LNLN channel path uses the same 3-bit integer serialization and decoding interface for its continuous features before channel transmission.

Gradient propagation through spiking neurons and discrete channels. The LIF neurons in the DMDN sender use the ATan surrogate-gradient function provided by SpikingJelly (Fang et al., 2023). During backpropagation, the thresholding operation is replaced by the derivative of the ATan surrogate function. The 3-bit integer conversion and bit serialization are discrete forward operations; the channel masks are sampled in the forward pass and treated as fixed in the backward pass. The BSC implements $\tilde{x} = (1-n)x + n(1-x)$; the BEC implements $\tilde{x} = (1-z)x + ze$ with $e = 0.5$. The LNLN path uses the same 3-bit conversion interface, with `de2bi` and `bi2de` implementing the forward quantization and decoding operations.

The results on the MOSI dataset are summarized in Table 4. We observe that: **Mild joint corruption can even regularize training:** At noise level $p = 0.1$, the model achieves an Acc-5 of 38.48, which is higher than the 37.46 obtained under the standard setting (i.e., $p = 0.0$), suggesting a slight beneficial regularization effect. **Graceful degradation under severe noise:** When $p \geq 0.3$, performance decreases but does not collapse abruptly, indicating strong robustness against heterogeneous perturbations. **Stable regression quality:** The MAE remains within a moderate range from 1.08 to 1.18, and the correlation values decrease smoothly rather than collapsing, suggesting that the regression behavior remains relatively stable under compounded distortions.

p	Acc-7	Acc-5	Acc-2	F1	MAE	Corr.
0.0	33.09	37.46	74.24 / 73.18	74.44 / 73.44	1.081	0.5190
0.1	34.11	38.48	72.87 / 71.72	72.83 / 71.77	1.091	0.5095
0.2	32.80	36.44	71.96 / 70.41	71.65 / 70.82	1.119	0.4913
0.3	32.07	36.44	72.41 / 72.39	71.43 / 71.39	1.132	0.4704
0.4	31.78	35.13	71.95 / 70.92	71.93 / 70.99	1.181	0.4269

Table 4: Performance under mixed BSC & BEC noise on MOSI dataset.

A.2 ADDITIONAL EXPERIMENTS AND ANALYSIS

This appendix provides supplementary experimental results and analyses omitted from the main paper due to space constraints.

A.2.1 SENSITIVITY ANALYSIS OF HYPERPARAMETERS

To further examine the robustness of our framework, we conducted a sensitivity analysis on the four loss weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, which correspond to the prediction, reconstruction, adversarial,

λ_1	λ_2	λ_3	λ_4	Acc-7	Acc-5	F1	Corr.
0.9	0.8	0.1	1.0	34.40	38.48	74.49 / 72.59	0.5190
0.0	0.8	0.1	1.0	32.65	35.71	72.15 / 70.98	0.4858
0.5	0.8	0.1	1.0	33.24	36.30	71.14 / 71.16	0.4926
1.0	0.8	0.1	1.0	33.97	37.32	71.73 / 72.10	0.5151
0.9	0.0	0.1	1.0	32.51	36.15	72.59 / 72.52	0.5095
0.9	0.5	0.1	1.0	33.09	38.05	71.91 / 72.01	0.5075
0.9	1.0	0.1	1.0	34.40	38.34	71.22 / 71.28	0.5021
0.9	0.8	0.0	1.0	30.61	33.97	71.80 / 70.70	0.4991
0.9	0.8	0.2	1.0	32.36	36.01	70.77 / 70.70	0.4896

Table 5: Sensitivity analysis of hyperparameters on MOSI.

and weight estimation objectives, respectively. In our main experiments, we followed LNLN and set $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.9, 0.8, 0.1, 1.0)$ to ensure a fair and consistent comparison, since both methods adopt a multi-task loss structure. However, it is important to assess how these hyperparameters influence performance.

To this end, we varied one hyperparameter at a time while keeping the others fixed, and evaluated the model on the MOSI dataset under BSC noise ($p = 0.2$). The results are summarized in Table 5. Several key insights emerge from the ablation results. First, the prediction loss λ_1 proves to be indispensable. Eliminating this term results in a noticeable performance drop: Acc-7 declines from 34.40 to 32.65, while the F1 score decreases from 72.59 to 70.98, confirming its role as the primary driver of task performance. The reconstruction loss λ_2 also plays a critical role, particularly in maintaining robustness under modality corruption. Without it, both Acc-7 and F1 are reduced, highlighting the importance of reconstruction in preserving cross-modal alignment. In contrast, the adversarial loss λ_3 has a more nuanced effect: moderate variations cause only minor differences, but completely discarding it results in a significant drop with Acc-7 decreasing to 30.61 and F1 to 70.70. This suggests that adversarial learning enhances the stability of latent representations and prevents overfitting. The weight estimation loss λ_4 is fixed at 1.0 in all rows, so this table does not evaluate its removal. Under the tested settings, λ_1 and λ_2 show the strongest effects, while the tested λ_3 values produce complementary variations across metrics.

Overall, λ_1 and λ_2 have the strongest observed influence under this sensitivity protocol, while the tested λ_3 settings provide complementary effects. This analysis supports the adopted hyperparameter setting and indicates that the auxiliary objectives contribute to the robustness of the framework.

A.2.2 Analysis of Model Stability

As shown in Table 6, both models experience gradual performance degradation as the random missing rate increases. LNLN performs better under mild corruption (missing rate ≤ 0.4), reflecting its advantage when sufficient modality information is preserved. In contrast, DMDN_{BSC} shows clearer advantages in Acc-2 and F1 when the missing rate becomes severe, especially at missing rates of 0.9 and 1.0. These results indicate that DMDN is not uniformly superior under mild missing conditions, where LNLN remains competitive or better, but the proposed dual-noise training is more beneficial in highly incomplete scenarios.

A.2.3 Ablation on Dominance Selection

To simulate scenarios where the model cannot adaptively select the dominant modality, we evaluate DIMC under three fixed-dominance configurations, in which text, audio, or vision is always designated as the dominant anchor regardless of its per-sample degradation. All variants share the full DMDN architecture and are evaluated on MOSI under a missing rate of 0.5 with BSC noise (BER $= 0.2$). As shown in Table 7, performance fluctuates with the chosen dominant modality, while the full DMDN generally outperforms all fixed-modality variants by dynamically adapting to input conditions: fixing text or audio preserves competitive fine-grained accuracy but weakens binary classification and regression quality, while fixing vision attains comparable Acc-7 but the lowest Acc-2 and F1 among all variants. In contrast, the full model with dynamic dominance se-

Missing Rate	Model	Acc-7	Acc-5	Acc-2	F1
0.1	LNLN	40.23	44.61	80.03	79.98
	DMDN_{BSC}	37.03	40.09	80.03	80.25
0.2	LNLN	36.30	41.25	78.43	78.44
	DMDN_{BSC}	36.30	39.50	78.05	78.21
0.3	LNLN	36.44	40.38	76.09	76.12
	DMDN_{BSC}	34.40	38.05	76.68	76.84
0.4	LNLN	36.30	38.05	74.64	74.66
	DMDN_{BSC}	31.34	37.17	75.15	75.38
0.5	LNLN	33.53	37.03	72.99	72.91
	DMDN_{BSC}	34.40	38.48	72.45	72.59
0.6	LNLN	28.13	33.67	68.95	69.01
	DMDN_{BSC}	28.72	31.63	69.05	69.27
0.7	LNLN	26.53	29.59	62.10	62.20
	DMDN_{BSC}	25.07	27.84	65.70	65.80
0.8	LNLN	23.47	27.70	59.33	59.43
	DMDN_{BSC}	22.74	27.11	60.67	60.69
0.9	LNLN	22.16	22.98	54.08	54.09
	DMDN_{BSC}	22.16	24.78	58.23	58.60
1.0	LNLN	18.08	17.64	51.75	51.86
	DMDN_{BSC}	17.35	17.94	55.95	60.88

Table 6: Performance comparison of random missing rates on MOSI.

lection achieves the best Acc-5, Acc-2, F1, MAE, and correlation. These results support our core claim that dynamically identifying and adapting to modality degradation leads to more robust and generalizable performance, especially when no single modality is consistently dominant.

Method	Acc-7 ↑	Acc-5 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
DIMC w/ fixed text dominance	33.09	36.15	71.87	72.17	1.1002	0.4921
DIMC w/ fixed audio dominance	33.97	37.90	71.87	71.90	1.0848	0.5094
DIMC w/ fixed vision dominance	**34.40**	37.76	71.14	71.48	1.0838	0.5175
DMDN_{BSC} (full)	**34.40**	**38.48**	**72.45**	**72.59**	**1.0670**	**0.5190**

Table 7: Ablation of dynamic dominance selection on MOSI under a missing rate of 0.5 with BSC noise.

A.2.4 Analysis of Computational Complexity

We provide a comparison of computational complexity and runtime with LNLN, a representative baseline with similar loss structure and task scope. As shown in Table 8, DMDN achieves a substantially lower FLOPs count, requiring only 314.19G compared with the 622.91G of the baseline. This corresponds to roughly a 50% reduction, largely attributable to the efficiency of the SNN and SSM modules. Although the average inference time increases from 0.078s to 0.103s, the runtime remains within a comparable range, and the additional cost mainly stems from the simulation of spiking dynamics. Importantly, the comparable parameter size and reduced FLOPs demonstrate that our method is computationally efficient and scalable, particularly given its added robustness under dual-noise conditions.

A.2.5 Analysis of Lazy Behavior

To examine whether DMDN exhibits *lazy behavior* similar to prior methods, we visualized the confusion matrices under varying random missing rates on the MOSI dataset. The true labels (from top to bottom) and predicted labels (from left to right) follow the sequence Strong Neg, Weak Neg, Neg, Neutral, Weak Pos, Pos, Strong Pos. Following the approach in LNLN, this analysis evalu-

Method	FLOPs (G)	Params. (M)	Avg. Time (s)
DMDN (Ours)	314.187	92.290	0.1027
LNLN	622.908	91.990	0.0783

Table 8: Theoretical operation-count estimates and measured inference time. The FLOPs counts include the stated network dimensions and $T_{\mathrm{snn}} = 8$ SNN simulation steps; timing was measured on an NVIDIA 4090D with batch size 64.

Hardware note. On dedicated neuromorphic hardware, event-driven execution can better realize the operation and energy advantages of sparse spikes; on a conventional GPU, sequential simulation across multiple timesteps introduces additional execution overhead.

ates whether the model tends to over-predict certain categories as the input information becomes increasingly incomplete.

As shown in Tables 9 and 10, and further visualized in Fig. 3, the predictions of DMDN remain distributed across multiple sentiment classes even at high missing rates. At a 50% missing rate, the model still generates outputs spanning intermediate classes T2–T5, without collapsing entirely to the central class T3. When the missing rate reaches 90%, although predictions concentrate more around T3 due to reduced modality information, there is no domination by a single class, indicating that the model continues to leverage available information to make nuanced decisions. These results contrast with the typical lazy behavior observed in less robust models, where predictions collapse to majority or central classes as the missing rate increases. The maintained diversity of predicted categories suggests that DMDN preserves dynamic decision-making and does not rely solely on biasing toward dominant classes when faced with incomplete inputs. This demonstrates the model's robustness and its capacity to extract useful knowledge even under severe modality corruption.

	P0	P1	P2	P3	P4	P5	P6
T0	0.106383	0.489362	0.170213	0.170213	0.021277	0.042553	0.000000
T1	0.077419	0.374194	0.187097	0.174194	0.096774	0.090323	0.000000
T2	0.054054	0.391892	0.277027	0.148649	0.094595	0.033784	0.000000
T3	0.019417	0.165049	0.300971	0.320388	0.126214	0.067961	0.000000
T4	0.000000	0.105263	0.210526	0.175439	0.166667	0.342105	0.000000
T5	0.010101	0.090909	0.141414	0.141414	0.232323	0.363636	0.020202
T6	0.050000	0.050000	0.150000	0.200000	0.100000	0.400000	0.050000

Table 9: Confusion matrix under 50% random missing rate on MOSI.

	P0	P1	P2	P3	P4	P5	P6
T0	0.042553	0.063830	0.191489	0.382979	0.276596	0.042553	0.000000
T1	0.006452	0.129032	0.187097	0.354839	0.258065	0.064516	0.000000
T2	0.000000	0.114865	0.256757	0.412162	0.135135	0.081081	0.000000
T3	0.000000	0.067961	0.184466	0.504854	0.174757	0.067961	0.000000
T4	0.000000	0.017544	0.114035	0.438596	0.236842	0.192982	0.000000
T5	0.000000	0.040404	0.121212	0.404040	0.212121	0.222222	0.000000
T6	0.000000	0.000000	0.050000	0.600000	0.300000	0.050000	0.000000

Table 10: Confusion matrix under 90% random missing rate on MOSI.

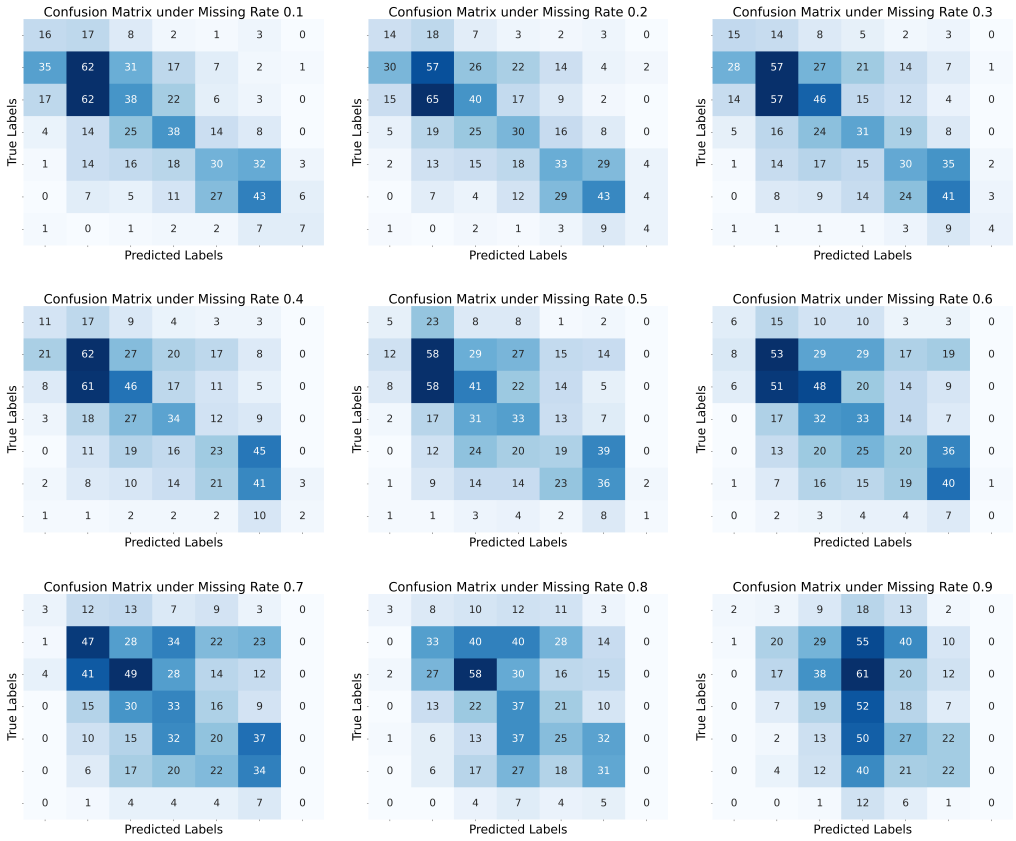

Figure 3: Confusion Matrices of DMDN on MOSI across Random Missing Rates. The true labels (from top to bottom) and predicted labels (from left to right) follow the sequence Strong Neg, Weak Neg, Neg, Neutral, Weak Pos, Pos, Strong Pos.