BiMIN: Bilateral Modality Imagination Network for Sentiment Analysis under Modality Uncertainty

Xuanchao Lin
School of Computer Engineering and Science
Shanghai University
Shanghai, China
linxuanchao@shu.edu.cn

Junjie Peng*
School of Computer Engineering and Science
Shanghai University
Shanghai, China
jjie.peng@shu.edu.cn

***Abstract*—Human communication relies on the synergy of textual, acoustic, and visual cues to convey sentiment. While Multimodal Sentiment Analysis (MSA) has made significant strides in capturing these interactions, most existing models are predicated on the idealized assumption of complete data availability. In real-world scenarios, however, one or more modalities may be entirely absent, causing standard models to suffer severe performance degradation. To address this challenge, we propose the Bilateral Modality Imagination Network (BiMIN), a unified framework designed for robust sentiment analysis under modality uncertainty. BiMIN overcomes the limitations of existing approaches through two core innovations. First, we design a novel Bilateral Auto-Encoder featuring parallel Core and Detail branches. This structure effectively disentangles global semantics from fine-grained details, ensuring robust feature preservation. Second, to uniformly handle diverse absence scenarios, we incorporate an Absence-Gating Mechanism within a cascaded architecture designed for robust iterative imputation. This mechanism dynamically modulates the integration of original and imagined features based on real-time modality availability, preventing the hallucination of noise. Experimental results on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets demonstrate that BiMIN achieves superior stability and the highest average accuracy. Specifically, it delivers remarkable performance gains, exceeding baselines by 3.36% on MOSEI, 1.04% on MOSI, and 4.97% on SIMS, while maintaining competitive proficiency even in full-modality settings.**

***Index Terms*—Uncertainty estimation, Multimodal representation, Sentiment analysis, Unified model**

I. Introduction

HUMAN perception relies on diverse sensory information (text, sound, and vision) to understand the world. Similarly, Multimodal Sentiment Analysis (MSA) aims to assess sentiments [1] and intents [2] by integrating these modalities. However, unlike ideal laboratory settings, real-world scenarios are fraught with modality uncertainty. For instance, in telephonic communication, visual cues are unavailable; in cross-lingual interactions, textual semantics may be unintelligible. As a result, practical applications often face scenarios where one or more modalities are completely absent.

Most existing MSA models [3]–[6] are tailored for complete trimodal data. While some focus on specific bimodal [7]–[9] or unimodal [10]–[12] settings, these rigid architectures lack flexibility. A model optimized for trimodal input typically suffers a severe performance decline when degraded to bimodal or unimodal inference, as its predictive capability heavily relies on the presence of the specific modalities it was designed for.

To clarify the research scope, it is crucial to distinguish between *missing* modalities (partial, random loss, e.g., noise) and *absent* modalities (complete unavailability, e.g., sensor failure). Existing literature predominantly targets *missing* modalities using Translation-based [13] or Transformer-based [14] methods. While these approaches attempt feature reconstruction, they often face broken dependency chains or incur excessive computational costs when applied to severe absence scenarios. Regarding *absent* scenarios, current solutions typically fall short due to inherent design limitations. Some approaches [15], [16] rely on distilling knowledge from a pre-trained full-modality teacher to a partial student. These methods restrict the flexibility of independent modality training. Other approaches focus on mitigating noise via uncertainty estimation [17], [18]. Yet, they often struggle to calibrate confidence when the dominant text modality is entirely absent, leading to potential confidence collapse due to distribution shifts. Similarly, Auto-Encoder-based approaches [19], [20] predominantly rely on single-branch architectures that force heterogeneous data into a unified latent space. This design creates an information bottleneck, causing feature entanglement where subtle details are overshadowed by dominant semantics.

To address these limitations, we propose the **Bilateral Modality Imagination Network (BiMIN)**. Unlike static approaches, BiMIN introduces a dynamic framework with two innovations. **First**, to resolve the feature entanglement bottleneck, we design a **Bilateral Auto-Encoder** that explicitly **disentangles** global semantics from fine-grained details. **Second**, distinct from passive weighting, we incorporate an **Absence-Gating Mechanism** to dynamically regulate information flow. This ensures the model intelligently switches between original and imagined features based on availability, preventing noise hallucination.

Experimental results on three public datasets indicate that our model achieves superior average performance and stability, consistently outperforming benchmarks in overall metrics and robustness.

The primary contributions of this work are:

*Corresponding author

- We design a Bilateral Modality Imagination Network (BiMIN) aiming to address sentiment analysis under modality uncertainty with potential absence.
- We design a novel Bilateral Auto-Encoder with parallel Core and Detail branches, which effectively solves the feature entanglement problem by independently preserving global semantics and fine-grained details.
- We introduce an Absence-Gating Mechanism that dynamically regulates the fusion of original and imagined features based on real-time modality availability, ensuring robust performance whether modalities are present or absent.
- Comprehensive experiments demonstrate that BiMIN achieves state-of-the-art stability and accuracy, specifically outperforming baselines by significant margins in severe modality-absence scenarios.

II. Related Work

A. Sentiment Analysis under Modality Certainty

Early sentiment analysis research primarily focused on unimodal approaches, treating text, audio, or vision in isolation [10]–[12]. However, these methods often fail to capture the complexity of human expression. To address this, Multimodal Sentiment Analysis (MSA) emerged to integrate complementary information, evolving from bimodal pairs [7], [8] to sophisticated trimodal fusion strategies. Prominent works utilize diverse mechanisms such as cross-modal attention [4], [21], self-supervised learning [3], [22], and graph-based interactions [23], [24] to achieve state-of-the-art performance. However, these models heavily rely on the assumption of **modality completeness**. Designed for static, well-defined inputs, they lack the flexibility to handle real-world scenarios where modalities are uncertain or absent, necessitating a paradigm shift toward more robust architectures.

B. Sentiment Analysis under Modality Uncertainty

To tackle the challenge of data incompleteness, recent research falls into five primary methodological categories as follows.

1) Translation-based Models: Models like MCTN [13], CTFN [25], and MTMSA [26] map available modalities to missing ones but fail when the source modality itself is absent, breaking the translation chain, making them ineffective for severe absence cases.

2) Transformer-based Models: Models such as TFR-Net [14] reconstruct missing features via cross-modal attention, while MPMM [27] and MPLMM [28] utilize prompt learning. Despite their effectiveness, these models typically suffer from high computational complexity due to heavy Transformer backbones.

3) Knowledge Distillation(KD)-based Models: Models like CorrKD [16], and HRLF [29] transfer knowledge from a complete-modality teacher to a partial-modality student. Despite their success, they rigidly require a pre-trained full-modality teacher. This dependency limits training flexibility, whereas real-world scenarios often necessitate independent modality processing without joint supervision.

4) Uncertainty Estimation(UE)-based Models: Models like LNLN [30], UEFN [17] and P-RMF [18] mitigate noise by dynamically re-weighting modalities with estimated uncertainty. Nevertheless, their uncertainty estimators are typically conditioned on a dominant modality (text) or trained under uniform-random missing priors. Consequently, when the dominant cue is absent or the real missing pattern deviates from the training prior, the estimated confidence becomes poorly calibrated and exhibits systematic bias or over-confidence.

5) Auto-Encoders(AEs)-based Models: Models like CRA [31], MMIN [19], and If-MMIN [20] reconstruct missing views via latent distribution learning. However, these approaches leads to feature entanglement, where dominant semantics overshadow fine-grained sensory cues, creating a performance bottleneck.

Different from aforementioned approaches, our proposed **BiMIN** represents a paradigm shift from passive, dependent, or heavy architectures to an active, efficient, and disentangled framework. Unlike **Translation and KD methods** that rely on rigid dependency chains or full-modality teachers, BiMIN allows for independent modality processing, offering superior training flexibility. In contrast to computationally intensive **Transformer-based models**, BiMIN avoids heavy attention backbones. Instead, it achieves a lightweight design with significantly lower FLOPs while maintaining SOTA performance, striking an optimal balance between accuracy and efficiency. Diverging from **UE models** that merely re-weight uncertain modalities, BiMIN employs an **Absence-Gating Mechanism** to actively detect availability and switch between perception and imagination. Most crucially, compared to **single-branch AEs**, BiMIN overcomes the feature entanglement bottleneck. By employing a **Bilateral Imagination strategy** with parallel dual-pathways, it explicitly decouples coarse-grained semantics from fine-grained details. This holistic design offers a unified, robust, and efficient solution for diverse absence scenarios.

III. Method

In this section, we begin by defining the problem, followed by a description of our BiMIN model.

A. Task Setup

Given a multimodal dataset containing N samples, we represent the raw input of the i-th sample as $x_i = (x_i^t, x_i^v, x_i^a)$, corresponding to the textual, visual, and acoustic modalities, respectively. Here, the superscript $m \in \{t, v, a\}$ denotes the specific modality. The dataset is defined as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where y_i represents the ground-truth sentiment polarity. The primary objective of this work is to predict the sentiment label y_i based on x_i, specifically addressing robust performance in scenarios where one or more modalities within x_i are completely absent.

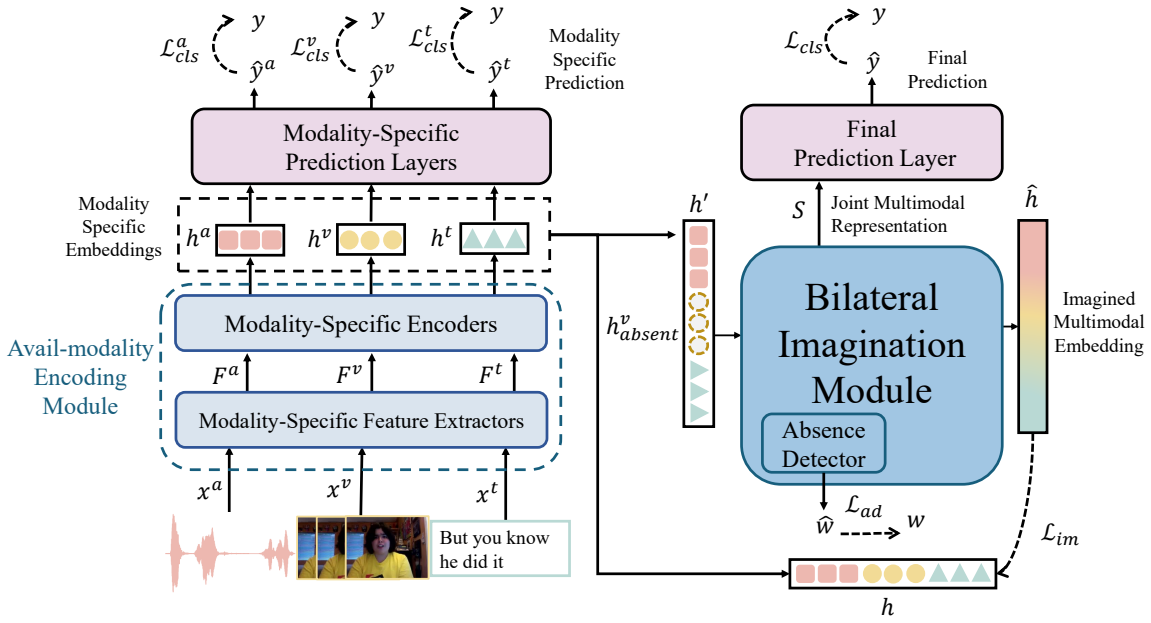

Fig. 1. Architecture of BiMIN (taking the visual modality absent condition as example). BiMIN is trained with all six possible absent modality conditions.

B. Overall Framework

The proposed BiMIN framework as illustrated in Fig. 1, comprises two primary components:

Avail-modality Encoding Module first leverages pre-trained models to extract high-level features from the raw data of each available modality. These features are subsequently mapped into a unified subspace via Modality-Specific Encoders to yield aligned embeddings h^m.

Bilateral Imagination Module cascades M Bilateral Auto-Encoders. It takes the initial concatenated embedding h' as input and iteratively refines it to generate the imagined embedding $\hat{h}$. Crucially, it aggregates the hierarchical hidden states S_i from each stage B_i ($i \in \{1, \ldots, M\}$) to form a robust joint multimodal representation S.

The framework adopts a stagewise training strategy. During the inference phase, the Modality Encoding Module processes the accessible raw inputs to produce h'. This embedding is then propagated through the Bilateral Imagination Module to synthesize the joint representation S, which is finally utilized by a prediction head to estimate the sentiment polarity $\hat{y}$.

C. Avail-modality Encoding Module

Modality-Specific Feature Extractors: Given a raw video, we extract unimodal sequences x^t (text), x^v (vision), and x^a (audio). We utilize pre-trained models to capture high-quality features: **BERT** [32] for textual features F^t, **X-CLIP** [33] for visual features F^v, and **Whisper** [34] for acoustic features F^a.

Modality-Specific Encoders: To align the extracted heterogeneous features into a unified space, we employ 1D convolutions to encode temporal information and a linear layer to project the representations into a unified dimension d. The resulting embeddings $h^m \in \mathbb{R}^d$ ($m \in \{t, v, a\}$) are concatenated to form the joint representation $h \in \mathbb{R}^{3d}$. To handle modality absence, we employ a **zero-masking strategy**: the embedding of any absent modality is substituted with a zero vector $\mathbf{0}_d$. For instance, if the visual modality is absent, the input is formatted as $h' = [h^t; \mathbf{0}_d; h^a] \in \mathbb{R}^{3d}$.

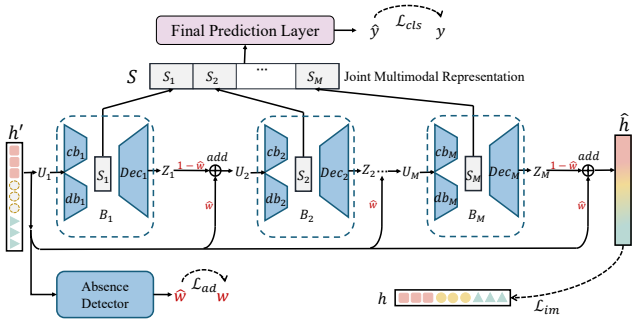

Fig. 2. Achitecture of Bilateral Imagination Module.

D. Bilateral Imagination Module

We propose a Bilateral Imagination Module to learn robust joint multimodal representations. This module cascades multiple Bilateral Auto-Encoders to progressively refine the representation quality.

1) Absence Detector: To dynamically adapt to modality availability, we employ an absence detector, denoted as E_{ad}, which consists of two linear layers followed by a classifier. E_{ad} takes the initial multimodal features h' as input and predicts a completeness mask $\hat{w}$. For instance, in a

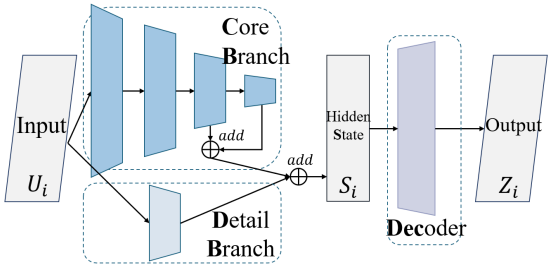

Fig. 3. Achitecture of Bilateral Auto-Encoder

trimodal setting where the feature dimension is d, if the visual modality is absent, the ground-truth label is defined as $w = (\mathbf{1}_d, \mathbf{0}_d, \mathbf{1}_d) \in \mathbb{R}^{3d}$. The predicted completeness mask $\hat{w}$ is obtained as follows:

$$\hat{w} = E_{ad}(h') \tag{1}$$

where $\hat{w} \in [0,1]^{3d}$ indicates the probability of modality presence.

2) Bilateral Auto-Encoder: The core requirement of our task is to generate high-quality hidden states that contribute to a robust multimodal representation while preventing feature loss. To achieve this, we design the **Bilateral Auto-Encoder** (Fig. 3). Unlike single-branch architectures that force heterogeneous information into a unified latent space, this module explicitly decouples the feature encoding process into two specialized parallel branches:

Detail Branch: This branch is designed to retain fine-grained features (e.g., specific sentiment fluctuations) that are often discarded by deep networks. Analogous to edge preservation in computer vision, we employ a shallow neural network with wide channels. Structurally, it functions as a residual connection to facilitate the direct flow of fine-grained information.

Core Branch: In parallel, this branch stacks multiple linear layers to capture deeper, abstract semantics. A global average pooling layer is applied at the end to incorporate global contextual information. By using a lower channel capacity, the Core Branch focuses on distilling the most essential semantics, while the Detail Branch acts as a complement.

These two branches jointly form the **encoder** of the Bilateral Auto-Encoder. Let $cb_i(\cdot)$ and $db_i(\cdot)$ denote the functions of the Core Branch and Detail Branch in the i-th layer, respectively. By physically separating the encoding pathways of global semantics and fine-grained details, this dual-path design effectively **disentangles** the representation. It ensures that the model can capture the global essence without losing the subtle nuances, thereby resolving the feature entanglement bottleneck.

For the **Decoder** (Dec_i), we adopt a simplified single-layer linear projection to balance reconstruction accuracy and computational efficiency.

3) Cascaded Residual Architecture: The Bilateral Imagination Module consists of M cascaded Bilateral Auto-Encoders $(B_1, B_2, ..., B_M)$. We introduce the Absence-Gating Mechanism controlled by the completeness mask $\hat{w}$ to regulate the input flow between stages. Let U_i, S_i, and Z_i denote the input, hidden state, and output of the i-th Bilateral Auto-Encoder, respectively. The computation flow is defined as:

$$U_i = \begin{cases} h' & \text{if } i = 1; \\ \hat{w} \odot h' + (1 - \hat{w}) \odot Z_{i-1}, & \text{if } i > 1. \end{cases} \tag{2}$$

$$S_i = cb_i(U_i) + db_i(U_i) \tag{3}$$

$$Z_i = Dec_i(S_i) \tag{4}$$

Here, $\odot$ denotes **element-wise multiplication**, and Eq. 2 represents the **Absence-Gating Mechanism**. When a modality is present ($\hat{w} \approx 1$), the term $\hat{w} \odot h'$ dominates, ensuring the model relies on the original, reliable embedding h'. Conversely, when a modality is absent ($\hat{w} \approx 0$), the term $(1 - \hat{w}) \odot Z_{i-1}$ takes over, forcing the module to impute the missing features using the "imagined" embedding Z_{i-1} from the previous stage. This allows for an iterative refinement of the imagined features across the cascade.

The final imagined multimodal embedding $\hat{h}$ is given by:

$$\hat{h} = BM(h') = \hat{w} \odot h' + (1 - \hat{w}) \odot Z_M \tag{5}$$

Where $BM(\cdot)$ represents the function of the Bilateral Imagination Module. We concatenate the hidden states S_i from each Bilateral Auto-Encoder B_i in the Bilateral Imagination Module to form a joint multimodal representation $S = (S_1; S_2; ...; S_M)$, which is used for subsequent sentiment analysis.

E. Prediction Layer

Finally, we utilize Multi-Layer Perceptrons (MLP) to perform sentiment regression. The modality-specific predictions and the final joint prediction are computed as $\hat{y}^m = \text{MLP}^m(h^m)$ and $\hat{y} = \text{MLP}(S)$, respectively, where h^m denotes the encoded features and S represents the joint multimodal representation.

F. Loss Function

In the process of BiMIN training, we adopt a stagewise training method.

In the Avail-modality Encoding Module, we train each modality-specific encoder within its respective modality. The training process of the unimodal initial embeddings is supervised by the classification loss $\mathcal{L}^m_{cls}$, which is defined as follows:

$$\mathcal{L}^m_{cls} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}^m - y| \tag{6}$$

Where $|\cdot|$ represents absolute error between ground-truth label y and modality-specific prediction $\hat{y}^m$.

In the Bilateral Imagination Module, three loss functions guide the training. First, the classification loss $\mathcal{L}_{cls}$ is used to supervise the training process with sentiment polarity targets. Secondly, the imagined loss $\mathcal{L}_{im}$ supervises the training of the

imagined multimodal embedding $\hat{h}$ with full-modality embeddings h. Thirdly, the absence detector loss $\mathcal{L}_{ad}$ supervises the training of the $\hat{w}$ with label of absence w:

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y} - y| \tag{7}$$

$$\mathcal{L}_{im} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{h} - h\|_2^2 \tag{8}$$

$$\mathcal{L}_{ad} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{w} - w\|_2^2 \tag{9}$$

Where $|\cdot|$ represents absolute error between ground-truth label y and final sentiment prediction $\hat{y}$, and $\|\cdot\|_2^2$ is the squared Frobenius norm. The overall loss $\mathcal{L}$ is calculated by adding these three functions together:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{im} + \beta\mathcal{L}_{ac} \tag{10}$$

Where α and β are hyperparameters.

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets: To replicate real-world conditions, we assess our model using CMU-MOSI [35], CMU-MOSEI [36], and CH-SIMS [37] that three well-established datasets for multimodal sentiment analysis.

2) Evaluation Metrics: MOSI and SIMS: Acc-2, F1 scores. MOSEI: Acc-2, F1 scores, Acc-5, Acc-7, MAE, Corr.

In the experimental results of the following sections: "Avg" represents the average performance across the six available modality combinations, while "Std" denotes the standard deviation of these performances. "↑" indicates that a higher value is better, whereas "↓" signifies that a lower value is preferable. Notably, for "Std", a lower value is desired. The best results are highlighted in **bold**, while the second-best results are underlined.

B. Experimental Setup

1) Training Set and Test Set: To simulate modality uncertainty, we derive the six possible absence combinations from the original full-modality dataset. For training, we construct a unified dataset by mixing samples from all six scenarios, enabling the model to learn robust representations across diverse conditions. For inference, we evaluate the model on six separate test subsets, each corresponding to a specific absence scenario. This allows us to rigorously assess the model's performance under each distinct absent pattern (e.g., evaluating inputs formatted as $(x^t, \mathbf{0}, x^a)$ specifically when vision is absent).

2) Model Training Details: In all experiments, we employ the Adam optimizer [38] with a batch size of 64. The model begins training with an initial learning rate of 1×10^{-3} using early stopping with patience of 8 epochs. To ensure reproducibility, we fix the random seed so that all models are trained on the same dataset. All models are constructed by employing the PyTorch deep learning framework.

C. Baselines

To evaluate the performance of BiMIN under modality uncertainty, we benchmark it against thirteen state-of-the-art models: MCTN(2019) [13], TFR-Net(2021) [14], MMIN(2021) [19], If-MMIN(2022) [20], MPMM(2023) [27], MPLMM(2024) [28], MTMSA(2024) [26], LNLN(2024) [30], UEFN(2024) [17], CorrKD(2024) [16], HRLF(2024) [29], FSRF(2025) [15], P-RMF(2025) [18].

Additionally, to verify the versatility of BiMIN in ideal complete-data scenarios, we compare it against these representative full-modality models: MulT(2019) [21], Self-MM(2021) [22], TETFN(2023) [3], ALMT(2023) [4], DMD(2023) [5], DLF(2025) [6], GsiT(2025) [1].

D. Experiments under Modality Uncertainty

To evaluate the robustness of BiMIN under modality uncertainty, we compared BiMIN against state-of-the-art baselines across three benchmark datasets (Table I).

1) Quantitative Analysis: BiMIN achieves the highest **Average Accuracy** on all datasets (MOSI: 77.82%, SIMS: 77.62%, MOSEI: 82.98%) and demonstrates exceptional stability with the lowest standard deviation on SIMS and MOSEI. Specifically, on MOSI, it outperforms the strongest baseline by $\sim$1%. On SIMS, it dominates every scenario, proving cross-lingual generalization. On the large-scale MOSEI, it leads by 3.9% in average accuracy with the lowest variance (Std 2.10), confirming its scalability.

2) Performance in Text-Absent Scenarios: In scenarios where the textual modality is absent ($\{a\}$, $\{v\}$, $\{a, v\}$), BiMIN demonstrates remarkable resilience, particularly on the larger MOSEI dataset. While some baselines specialize in specific bimodal combination ($\{a, v\}$), they often fail in single-modality scenarios ($\{a\}$, $\{v\}$). In contrast, BiMIN consistently dominates single-modality scenarios. For instance, on MOSEI, BiMIN surpasses the second-best model (P-RMF) by over **6%** in both $\{a\}$ (82.20% vs. 75.91%) and $\{v\}$ (80.02% vs. 73.19%) conditions. This confirms that our **Bilateral Auto-Encoder** successfully preserves fine-grained acoustic and visual cues that are otherwise lost in other architectures.

3) Performance in Text-Available Scenarios: When the textual modality is available ($\{t\}, \{a, t\}, \{v, t\}$), BiMIN remains highly competitive. Although specific task-oriented baselines achieve marginally higher scores in certain combinations, BiMIN consistently ranks within the top tier and often leads. Crucially, this stable performance validates the effectiveness of our **Absence-Gating Mechanism**. Unlike static fusion methods that might introduce artifacts regardless of input quality, our gating mechanism accurately detects the presence of the dominant textual modality. It ensures that the model prioritizes these high-quality original features and suppresses unnecessary imagination, thereby preserving the semantic integrity of the input without introducing noise.

4) Efficiency Analysis: As shown in Table II, we evaluate the trade-off between model complexity and performance. Compared to Transformer-based models (e.g., TFR-Net), BiMIN reduces FLOPs by an order of magnitude while

TABLE I

QUANTITATIVE RESULTS (%) OF 6 ABSENT-MODALITY COMBINATIONS. MODEL DATA SOURCES NOTED; UNLABELED ONES ARE CODE-REPRODUCED.

Datasets	Model	$\{a\}$		$\{v\}$		$\{t\}$		$\{a,v\}$		$\{v,t\}$		$\{a,t\}$		Avg		Std	
		Acc2↑	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2**	**F1**
MOSI	MCTN [13]	-	-	-	-	79.30	79.10	53.10	53.20	76.80	76.80	76.40	76.40	-	-	-	-
	TFR-Net [14]	59.45	59.49	48.69	40.16	82.32	82.37	59.76	60.00	82.93	**82.99**	81.86	81.83	69.17	67.81	13.70	15.99
	MMIN [19]	51.38	47.66	52.99	51.02	81.86	81.72	52.13	49.16	79.57	75.65	75.76	75.86	65.50	64.18	13.68	13.03
	If-MMIN [20]	62.04	60.93	53.35	51.38	81.10	81.06	59.30	50.04	79.27	79.39	77.74	77.88	68.80	66.78	10.92	13.15
	MPMM [28]	57.26	59.35	58.63	59.12	79.81	80.10	60.54	61.33	80.74	80.93	79.89	79.84	69.48	70.11	10.71	10.21
	MPLMM [28]	62.71	63.65	63.12	63.74	80.12	80.31	65.02	65.41	81.12	81.19	80.76	81.09	72.14	72.57	8.56	8.32
	MTMSA [26]	60.93	-	57.29	-	78.64	-	61.97	-	80.72	-	82.29	-	70.31	-	10.40	-
	LNLN [30]	-	-	-	-	-	-	55.79	63.10	72.31	73.67	72.32	73.68	-	-	-	-
	UEFN [17]	45.36	45.94	49.78	49.78	79.69	79.71	51.78	50.34	82.99	82.64	80.78	80.87	65.06	64.88	16.32	16.27
	CorrKD [16]	66.52	-	60.72	-	81.20	-	73.74	-	83.58	-	**84.73**	-	75.08	-	8.98	-
	HRLF [29]	69.47	-	64.59	-	83.36	-	75.62	-	83.56	-	83.86	-	76.74	-	7.56	-
	FSRF [15]	59.23	-	64.68	-	82.58	-	**76.25**	-	**83.72**	-	84.19	-	76.78	-	7.53	-
	P-RMF [18]	71.44	-	70.32	-	81.36	-	73.11	-	81.94	-	82.10	-	76.71	-	**5.16**	-
	Ours	**72.56**	**72.64**	**70.88**	**71.03**	**83.38**	**83.13**	73.63	**73.68**	82.93	82.74	83.54	**83.37**	**77.82**	**77.77**	5.52	**5.37**
SIMS	TFR-Net [14]	68.71	57.63	69.37	56.82	68.93	64.20	68.49	58.54	69.58	57.72	68.93	66.56	69.00	60.25	**0.37**	3.73
	MMIN [19]	69.58	62.98	68.93	57.00	77.68	77.39	66.74	63.34	75.05	75.24	77.24	76.32	72.54	68.71	4.29	7.90
	If-MMIN [20]	70.46	59.30	67.61	59.25	77.90	75.48	67.83	62.52	75.49	75.49	76.59	75.03	72.65	67.85	4.17	7.57
	MPMM [28]	64.98	76.41	65.40	77.92	78.56	78.56	64.01	73.47	77.51	77.47	77.11	77.20	71.26	76.85	6.49	1.64
	MPLMM [28]	65.93	**77.10**	66.02	**78.86**	**79.75**	78.74	65.28	74.02	77.97	77.95	77.45	77.84	72.07	77.42	6.37	**1.63**
	Ours	**72.65**	73.67	**75.93**	75.64	79.43	**79.26**	**74.84**	**75.75**	**83.02**	**83.25**	**79.87**	**80.15**	**77.62**	**77.95**	3.48	3.24
MOSEI	TFR-Net [14]	62.85	48.51	62.58	60.67	84.15	84.06	62.96	63.35	81.78	82.04	84.56	84.39	73.15	70.50	10.39	13.79
	MMIN [19]	57.95	58.66	59.60	60.26	82.97	83.03	64.42	64.63	83.90	83.92	82.20	82.45	71.84	72.16	11.36	11.13
	If-MMIN [20]	70.27	60.22	71.17	59.77	84.15	84.13	58.83	59.18	82.72	82.73	84.37	84.37	75.25	71.73	9.39	12.02
	MPMM [28]	66.94	68.74	67.21	69.27	78.21	78.30	68.11	69.79	79.63	79.71	79.41	79.47	73.25	74.17	5.86	4.98
	MPLMM [28]	67.33	68.71	67.29	69.40	79.12	79.17	68.21	69.91	80.11	80.13	80.45	80.43	73.75	74.98	6.16	5.31
	UEFN [17]	57.64	59.12	58.41	59.89	81.72	82.01	**82.84**	**82.41**	82.77	82.63	82.81	82.75	74.32	74.80	11.63	10.82
	CorrKD [16]	66.09	-	62.30	-	80.76	-	71.92	-	81.28	-	81.74	-	74.02	-	7.77	-
	HRLF [29]	69.32	-	64.90	-	82.05	-	73.80	-	81.09	-	82.62	-	76.74	-	6.81	-
	FSRF [15]	69.12	-	65.68	-	82.62	-	74.98	-	81.35	-	82.27	-	75.90	-	6.83	-
	P-RMF [18]	75.91	-	73.19	-	81.93	-	76.88	-	**85.17**	-	84.61	-	79.62	-	4.54	-
	Ours	**82.20**	**82.01**	**80.02**	**80.16**	**85.28**	**85.00**	80.75	79.96	84.76	**84.65**	**84.87**	**84.76**	**82.98**	**82.76**	**2.10**	**2.15**

TABLE II

TRAINABLE PARAMETERS (PARAMS$\times 10^6$) AND FLOATING POINTS OF OPERATIONS (FLOPS$\times 10^{10}$) OF MODELS WITH FOUR AUTO-ENCODERS.

Model	MOSI		SIMS		MOSEI	
	Params	**FLOPs**	**Params**	**FLOPs**	**Params**	**FLOPs**
MCTN	0.954	1.95	0.954	1.52	0.954	1.95
TRFNRT	96.9	61.2	90.8	45.4	99.9	63.2
MMIN	3.43	1.55	4.17	1.29	3.52	1.94
IF-MMIN	3.50	2.32	4.23	1.92	3.59	2.90
Ours	2.98	7.30	2.98	7.22	2.98	7.30

TABLE III

QUANTITATIVE RESULTS (%) ON MOSEI (FULL-MODALITY). MODEL DATA SOURCES NOTED.

Model	Acc2↑	F1↑	Acc5↑	Acc7↑	MAE↓	Corr↑
Mult [21]	82.50	82.30	-	51.80	0.580	0.703
Self-MM [22]	85.17	85.30	-	-	**0.530**	**0.765**
TETFN [3]	85.18	85.27	-	-	0.551	0.748
DMD [5]	84.80	84.70	-	54.60	-	-
DLF [6]	85.42	85.27	**55.70**	53.90	0.536	0.764
GsiT [1]	**85.60**	85.20	-	54.10	0.536	0.764
Ours	85.50	**85.42**	53.81	52.37	0.552	0.743

delivering superior accuracy. Although BiMIN incurs a slight increase in flops compared to simple AE-based models (like MMIN), this is a negligible cost for the substantial performance gains observed. The computational overhead remains within a manageable range, confirming that BiMIN strikes an optimal balance between accuracy and efficiency.

E. Experiments under Full-Modality Settings

To validate BiMIN as a versatile unified framework, we compared it against state-of-the-art models specialized for full-modality settings on the MOSEI dataset. As shown in Table III, BiMIN achieves exceptional performance, securing the **highest F1 score (85.42%)** and the **second-best Acc-2 (85.50%)** among the compared methods. It surpasses classic baselines like MulT and matches sophisticated models like Self-MM and GsiT. Although BiMIN is not explicitly optimized for fine-grained regression (Acc-7), its binary classification performance proves that the proposed architecture is highly effective in extracting and integrating multimodal features. This demonstrates that BiMIN serves as a robust unified backbone capable of adapting to diverse scenarios, delivering SOTA-level performance even when complete data is available.

TABLE IV
ABLATION STUDY RESULTS ON MOSEI. "AM" DENOTES AVAIL-MODALITY ENCODING MODULE; "BM" DENOTES THE BILATERAL IMAGINATION MODULE; "BIAE" DENOTES THE BILATERAL AUTO-ENCODER; "AD" DENOTES THE ABSENCE DETECTOR.

Model	$\{a\}$		$\{v\}$		$\{t\}$		$\{a,v\}$		$\{v,t\}$		$\{a,t\}$		Avg		Std	
	Acc2↑	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2↑**	**F1↑**	**Acc2**	**F1**
w/o AM	68.10	67.31	64.06	63.58	84.34	84.29	66.68	66.34	84.62	84.58	84.56	84.52	75.39	75.10	9.19	9.43
w/o AM&BM	57.95	58.66	59.60	60.26	82.97	83.03	64.42	64.63	83.90	83.92	82.20	82.45	71.84	72.16	11.36	11.13
w/o BM(BiAE&AD)	77.99	77.63	78.34	78.45	82.39	82.19	76.97	76.46	82.88	82.61	83.87	83.83	80.41	80.40	2.71	2.79
w/o BiAE	78.99	77.48	79.36	78.75	82.66	82.57	79.07	78.07	83.52	83.26	84.09	83.72	81.28	80.64	2.79	2.59
w/o AD	77.68	77.23	79.63	78.50	84.26	84.24	79.31	77.47	83.13	82.84	84.01	83.90	81.34	80.74	2.56	2.97
BiMIN	**82.20**	**82.01**	**80.02**	**80.16**	**85.28**	**85.00**	**80.75**	**79.96**	**84.76**	**84.65**	**84.87**	**84.76**	**82.98**	**82.76**	**2.10**	**2.15**

F. Ablation Experiments

To verify the contribution of each component within BiMIN, we conduct a comprehensive ablation study on the MOSEI dataset. The results are summarized in Table IV.

1) Impact of AM and BM: We first evaluate the fundamental contribution of the Avail-modality Encoding Module (AM) and the Bilateral Imagination Module (BM). Removing the AM (w/oAM) causes a significant performance drop (Avg Acc2: 75.39%), confirming that high-quality features provide a necessary foundation. However, high-quality features alone are insufficient. The performance of w/oAM (75.39%) is noticeably higher than the bare baseline $w/oAM\&BM$ (71.84%), but it still lags far behind the full BiMIN (82.92%). Crucially, removing the entire BM (w/oBM) results in a substantial decline to 80.41%. This 2.51% gap underscores that BM is not redundant; rather, it is essential for synthesizing robust multimodal representations from the provided features.

2) Superiority of Bilateral Auto-Encoder(BiAE): To isolate the benefit of our dual-branch design, we compare BiMIN with a single-branch variant ($w/oBiAE$). The results reveal that the single-branch design hits a performance ceiling (Avg Acc2: 81.28%), likely due to the information bottleneck where fine details are lost. In contrast, by introducing the dual-branch BiAE, BiMIN achieves an additional **1.64% improvement** (Avg Acc2: 82.92%). This empirical evidence strongly supports our hypothesis: the **Detail Branch** successfully preserves fine-grained cues that are otherwise compressed in standard architectures.

3) Necessity of Absence Detector(AD): Finally, we analyze the dynamic gating mechanism by comparing BiMIN with a static fusion variant (w/oAD). Removing the Absence Detector leads to a clear decline in overall accuracy (81.34%) and stability. This demonstrates that the **Absence-Gating Mechanism** is vital for intelligently regulating the ratio of original versus imagined features, preventing the model from hallucinating noise when valid data is present.

The ablation study confirms a clear hierarchy of contributions: while AM provides high-quality input, the Bilateral Imagination Module (BM) and its sub-components (BiAE and AD) are indispensable for maximizing performance. The full synergy of these components allows BiMIN to achieve the highest accuracy (82.98%) and stability (Std 2.10).

V. CONCLUSION

In this paper, we present the **Bilateral Modality Imagination Network (BiMIN)**, a unified framework designed to address the critical challenge of sentiment analysis under modality uncertainty, particularly in scenarios with absent modalities. BiMIN distinguishes itself through two core innovations: the **Bilateral Auto-Encoder** and the **Absence-Gating Mechanism**. The Bilateral Auto-Encoder overcomes the information loss inherent in single-branch architectures by employing parallel paths to simultaneously preserve coarse-grained semantic consistency and fine-grained detailed variations. Complementing this, the Absence-Gating Mechanism dynamically detects modality availability and regulates the information flow within a cascaded architecture. This ensures that the model intelligently balances the reliance on reliable observed data versus generated imagined features, rather than using a static fusion strategy.

Extensive experiments on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets demonstrate that BiMIN consistently outperforms state-of-the-art baselines across diverse absent-modality combinations, exhibiting superior stability and computational efficiency.

In future work, we aim to explore end-to-end fine-tuning strategies that jointly optimize the feature extraction and imagination modules, potentially further enhancing the model's adaptability and performance in complex real-world environments.

REFERENCES

[1] Y. Jin, J. Peng, X. Lin *et al.*, "Multimodal transformers are hierarchical modal-wise heterogeneous graphs," in *The 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[2] Z. Li, J. Peng, X. Lin, and Z. Cai, "Multimodal intent recognition based on text-guided cross-modal attention," *Applied Intelligence*, vol. 55, pp. 1–14, 2025.

[3] D. Wang, X. Guo, Y. Tian *et al.*, "Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

[4] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. 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*, 2023, pp. 756–767.

[5] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6631–6640.

[6] P. Wang, Q. Zhou, Y. Wu *et al.*, "Dlf: Disentangled-language-focused multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 20, 2025, pp. 21 180–21 188.

[7] S. Liu, W. Quan, Y. Liu, and D.-M. Yan, "Bi-directional modality fusion network for audio-visual event localization," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4868–4872.

[8] Y. Lin, P. Ji, X. Chen, and Z. He, "Lifelong text-audio sentiment analysis learning," *Neural Networks*, vol. 162, pp. 162–174, 2023.

[9] C. Li, J. Wang, H. Wang, M. Zhao, W. Li, and X. Deng, "Visual-texual emotion analysis with deep coupled video and danmu neural networks," *IEEE Transactions on Multimedia*, vol. 22, pp. 1634–1646, 2019.

[10] D. Tiwari and B. Nagpal, "Keaht: A knowledge-enriched attention-based hybrid transformer model for social sentiment analysis," *New Generation Computing*, vol. 40, pp. 1165–1202, 2022.

[11] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert Systems with Applications*, vol. 214, p. 118943, 2023.

[12] Y. Liu, C. Feng, X. Yuan *et al.*, "Clip-aware expressive feature learning for video-based facial expression recognition," *Information Sciences*, vol. 598, pp. 182–195, 2022.

[13] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 6892–6899.

[14] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4400–4407.

[15] Z. Liu, P. Chu, S. Dong *et al.*, "Fsrf: Factorization-guided semantic recovery for incomplete multimodal sentiment analysis," 2025.

[16] M. Li, D. Yang, X. Zhao *et al.*, "Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 12 458–12 468.

[17] S. Wang, K. Ratnavelu, and A. S. Bin Shibghatullah, "Uefn: Efficient uncertainty estimation fusion network for reliable multimodal sentiment analysis: Uefn: Efficient uncertainty estimation..." *Applied Intelligence*, vol. 55, no. 3, 2024.

[18] A. Zhu, M. Hu, X. Wang *et al.*, "Proxy-driven robust multimodal sentiment analysis with incomplete data," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 22 123–22 138.

[19] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 2608–2618.

[20] H. Zuo, R. Liu, J. Zhao, G. Gao, and H. Li, "Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[22] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 10 790–10 797.

[23] T. Zhao, J. Peng *et al.*, "A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis," *Appl. Intell.*, vol. 53, pp. 30 455–30 468, 2023.

[24] L. Wang, J. Peng *et al.*, "A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning," *Information Processing and Management*, vol. 61, p. 103675, 2024.

[25] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 5301–5311.

[26] Z. Liu, B. Zhou, D. Chu *et al.*, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, vol. 101, p. 101973, 2024.

[27] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee, "Multimodal prompting with missing modalities for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 943–14 952.

[28] Z. Guo, T. Jin, and Z. Zhao, "Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 1726–1736.

[29] M. Li, D. Yang, Y. Liu *et al.*, "Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 28 515–28 536.

[30] H. Zhang, W. Wang, and T. Yu, "Towards robust multimodal sentiment analysis with incomplete data," *arXiv preprint arXiv:2409.20012*, 2024.

[31] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1405–1414.

[32] J. Devlin, M. Chang, K. Lee, and K. 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: Human Language Technologies*, 2019, pp. 4171–4186.

[33] B. Ni, H. Peng, M. Chen *et al.*, "Expanding language-image pretrained models for general video recognition," in *European Conference on Computer Vision*, 2022, pp. 1–18.

[34] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, 2023, pp. 28 492–28 518.

[35] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, pp. 82–88, 2016.

[36] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. 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, pp. 2236–2246.

[37] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. 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*, 2020, pp. 3718–3727.

[38] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations.*, 2015, pp. 1–15.