Two Challenges, One Solution: Robust Multimodal Learning through Dynamic Modality Recognition and Enhancement

Lanxin Bi[1], Yunqi Zhang[1], Luyi Wang[1], Yake Niu[1], Hui Zhao[1,2*]

[1]Software Engineering Institute, East China Normal University
[2]Shanghai Key Laboratory of Trustworthy Computing, Shanghai, China
lanxin.bi@stu.ecnu.edu.cn,{yunqi.zhang, luyi.wang, yake.niu}@stu.ecnu.edu.cn,
hzhao@sei.ecnu.edu.cn

Abstract

Multimodal machine learning is often hindered by two critical challenges: modality missingness and modality imbalance. These challenges significantly degrade the performance of multimodal models. The majority of existing methods either require the availability of full-modality data during the training phase or necessitate explicit annotations to detect missing modalities. These dependencies severely limit the models' applicability in the real world. To tackle these problems, we propose a Dynamic Modality Recognition and Enhancement for Adaptive Multimodal fusion framework (*DREAM*). Within DREAM, we innovatively employ a sample-level dynamic modality assessment mechanism to direct selective reconstruction of missing or underperforming modalities. Additionally, we introduce a soft masking fusion strategy that adaptively integrates different modalities according to their estimated contributions, enabling more accurate and robust predictions. Experimental results on three benchmark datasets consistently demonstrate that DREAM outperforms several representative baseline and state-of-the-art models, marking its robustness against modality missingness and imbalanced modality.

1 Introduction

Multimodal machine learning, inspired by humans' ability to solve problems using information from varying modalities, such as acoustic, visual, and textual cues, focuses on enabling models to effectively leverage multimodal data (Liang et al., 2024). It has been successfully applied across various domains, including medical (Zhang et al., 2024b; Yao et al., 2024), public safety (Zhao et al., 2024), and multimodal sentiment analysis (Sun et al., 2022; Han et al., 2021).

Despite its success, multimodal machine learning often faces the challenge of missing modalities in real-world data, caused by sensor failures, data corruption, or privacy constraints. Modality missingness is random and prevalent, impacting both the training and inference phases. In real-world scenarios, models built on the assumption that all modalities are fully available (Yu et al., 2022; Sun et al.; Li et al., 2023; Tsai et al., 2019) may be misguided by missing modalities (Ma et al., 2022), as shown in Figure 1. It suggests that those models are sensitive to missing modalities.

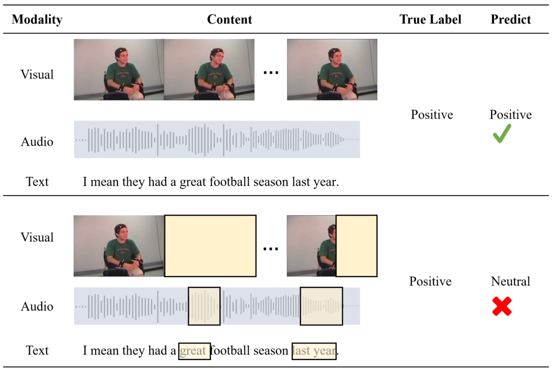
Modality
Content
True Label
Predict
Visual
Audio
Text
I mean they had a great football season last year.
Positive
Positive
✓
Visual
Audio
Text
I mean they had a great football season last year.
Positive
Neutral
✗

Figure 1: The yellow masks indicate the missing modalities at these positions. Without specific designs to handle modality missingness, models may predict 'Happy' when all modalities are present, but shift to 'Neutral' when some modalities are missing.

Although a wide range of methods have been proposed to address modality missingness (Zeng et al., 2022a; Li et al., 2024a; Guo et al., 2024; Li et al., 2024b), many of them face practical limitations. These approaches generally rely on additional information to function effectively. Some approaches (Li et al., 2023; Hu et al., 2020; Wang et al., 2023; Li et al., 2024b) assume the availability of complete multimodal data during the training phase. However, such an assumption is often unrealistic in real-world scenarios. Other methods depend on explicit annotations that indicate which modalities are missing for each input sample (Zeng

[*]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12855–12867
November 4-9, 2025 ©2025 Association for Computational Linguistics

et al., 2022a; Sun et al., 2024; Guo et al., 2024). While these annotations can guide the learning process, they impose a substantial manual workload, demanding consistent and accurate labeling across extensive datasets. Moreover, these annotations may not always be reliable or even feasible in dynamic, open-world settings. Therefore, these limitations make existing solutions difficult to scale or generalize.

Moreover, current methods for addressing modality missingness primarily focus on alleviating the performance decline resulting from incomplete data (Lian et al., 2023; Zhao et al., 2021; Ma et al., 2021). However, they neglect a more profound and inherent issue in multimodal machine learning: modality imbalance. Modality imbalance refers to the unequal learning proficiency of a model across different modalities, where the model may over-rely on certain modalities while under-utilizing or mislearning others. In such scenarios, dominant modalities disproportionately influence predictions, whereas other modalities contribute minimally or even introduce noise(Du et al., 2023; Wang et al., 2020a; Peng et al., 2022). Modality missingness and modality imbalance are two distinct problems, yet they often co-occur and may mutually reinforce each other. Modality imbalance can intensify the impact of missing modalities. It makes models more vulnerable when dominant modalities are absent. Conversely, missing modalities can exacerbate the imbalance by further restricting the learning of modalities with lower contributions.

In this work, we propose **D**ynamic Modality **R**ecognition and **E**nhancement for **A**daptive **M**ultimodal Fusion framework (*DREAM*), a novel framework that addresses both missing modality and imbalanced modality dynamically without explicit labels. Notably, DREAM does not require full-modality data during training or manual annotations to identify missing modalities like previous works do.

Specifically, DREAM dynamically estimates the actual contribution of each modality. Based on these estimates, DREAM identifies modalities that are missing or contribute little to the prediction and selectively reconstructs and enhances them. To effectively integrate both reconstructed and observed modalities, we introduce an adaptive fusion mechanism in DREAM. By dynamically modulating the fusion process according to the estimated modality contributions, the framework produces more effective and resilient decisions, even when dealing with missing or imbalanced modalities.

Our main contributions are summarized as follows:

- We propose **DREAM**, a robust multimodal framework that effectively addresses the challenges of missing and imbalanced modalities simultaneously via dynamic contribution estimation and enhancement.
- We design a dynamic modality evaluation mechanism that estimates each modality's contribution without relying on missing labels or full-modality supervision.
- Experimental results on three benchmark datasets (IEMOCAP, CMU-MOSI, and CMU-MOSEI) show that DREAM consistently outperforms baseline and state-of-the-art models under both incomplete and imbalanced modality conditions.

2 Related Work

Missing Modality Issue. In response to the challenge of modality missingness, previous research has primarily explored two major directions: joint representation methods and generative methods.

Joint representation methods aim to align the joint representations of samples with missing modalities to those of complete samples. These methods can approximate a unified multimodal embedding even in the absence of certain modalities (Zeng et al., 2022a; Li et al., 2024a,b; Liaqat et al., 2025; Ganhör et al., 2024). For instance, Zhao et al. (2021) leverage Cycle Consistency Learning to learn and predict robust joint multimodal representations from available modalities under uncertain missing-modality conditions. However, these methods generally require the availability of full-modality data during training for distillation learning, limiting their applicability in scenarios with missing modalities in training data.

Meanwhile, generative methods focus on reconstructing the missing modalities from the available ones to restore complete multimodal inputs for downstream tasks (Tran et al., 2017; Zeng et al., 2022b). Such approaches often leverage generative models, such as autoencoders (AEs) (Baldi, 2012) or generative adversarial networks (GANs) (Zhang et al., 2019), to restore the missing data. For example, Ma et al. (2021) propose a model named SMIL, adopting Bayesian meta-learning to cope with severe modality missingness. These methods usually require special annotations to mark the missing modalities. This additional annotation process not only increases the complexity of data preprocessing but also may introduce errors or biases, especially when dealing with large-scale datasets.

More recently, the remarkable progress in pre-trained models, particularly large language models (LLMs), has spurred the development of prompt-based techniques for modality missingness (Lee et al., 2023). These methods adapt pre-trained models through lightweight prompting strategies, enabling them to handle missing modalities without requiring full model retraining. Guo et al. (2024) utilizes prompt-based techniques to provide the pre-trained multimodal model with missing modality information at different stages, which empowers the model to effectively handle modality missingness. Meanwhile, Kim and Kim (2024) employ prompt-based techniques to integrate multiple high-performing pre-trained unimodal models to handle the missing modalities issue. While promising, many prompt-based methods still assume access to training data with full modalities or rely on predefined missing-modality indicators.

Imbalanced Modality Issue. Modal imbalance refers to the insufficient learning of individual modalities during the joint training of multimodal systems (Du et al., 2023; Wang et al., 2020a; Huang et al., 2022; Wu et al., 2022). This is often caused by imbalanced contributions from different modalities. Multimodal machine learning models tend to over-rely on the dominant modality (most commonly text) while neglecting meaningful learning from other inputs (such as audio or vision). Moreover, the dominant modality is also not fully learned because of the complementary information of other modalities. Thus, the under-utilization of modalities degrades the overall performance and harms the generalization ability of the model.

Recent efforts have explored several directions to mitigate this issue. Peng et al. (2022) controls the optimization of each modality by monitoring the discrepancy of their contribution to the learning objective. Zhang et al. (2024a) employ an alternating training scheme that separately learns unimodal features and a shared global representation, ensuring that the model captures both individual modality characteristics and cross-modal interactions.

Importantly, modality missingness can further aggravate modality imbalance, as the absence of one or more modalities forces the model to depend even more heavily on the dominant ones, making it harder to learn from underrepresented modalities.

3 Method

3.1 Problem Definition and Notation

We consider a multimodal dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{\mathcal{N}}$, consisting of $\mathcal{N}$ samples. Each sample $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_M^{(i)}\}$ contains M modalities, and $y^{(i)} \in \mathcal{Y}$ denotes the corresponding label. To simplify notation, we denote a generic sample as $x = \{x_1, x_2, \ldots, x_M\}$, where x_m represents the input from the m-th modality and $m \in \mathcal{M} = \{1, 2, \ldots, M\}$. We denote the missing modality as $\overline{x_m}$. The objective is to accurately predict the label y based on the given input x, even when certain modalities $\overline{x_m}$ are missing.

3.2 Overall Architecture

DREAM consists of three key modules: the Dynamic Modality Contribution Evaluation Module, the Modality Enhancement Module, and the Modality-aware Masked Fusion Module. Figure 2 illustrates the overall architecture of DREAM.

The Dynamic Modality Contribution Evaluation Module estimates the contribution of each modality using two complementary metrics: Predictive Accuracy Score (PAS) and Prediction Shift Score (PSS). Based on these assessments, the Modality Enhancement Module utilizes the PSS to selectively reconstruct missing or low-contribution modalities. The resulting enhanced modalities, along with the original available modalities, are then integrated by the Modality-aware Masked Fusion Module, which employs PSS as a soft mask to guide adaptive fusion. To ensure that the fusion strategy reflects the predictive utility of each modality, an auxiliary objective is employed to align the learned fusion weights with the PAS values.

3.3 Dynamic Modality Contribution Evaluation Module

The Dynamic Modality Contribution Evaluation Module leverages a pre-trained model to assess the contribution of each modality to the final prediction. Specifically, this module computes two metrics: PSS and PAS. Given that the pattern of modality missingness can vary among samples, this module performs sample-level evaluation. We apply Shapley value (Weber, 1988) to compute the

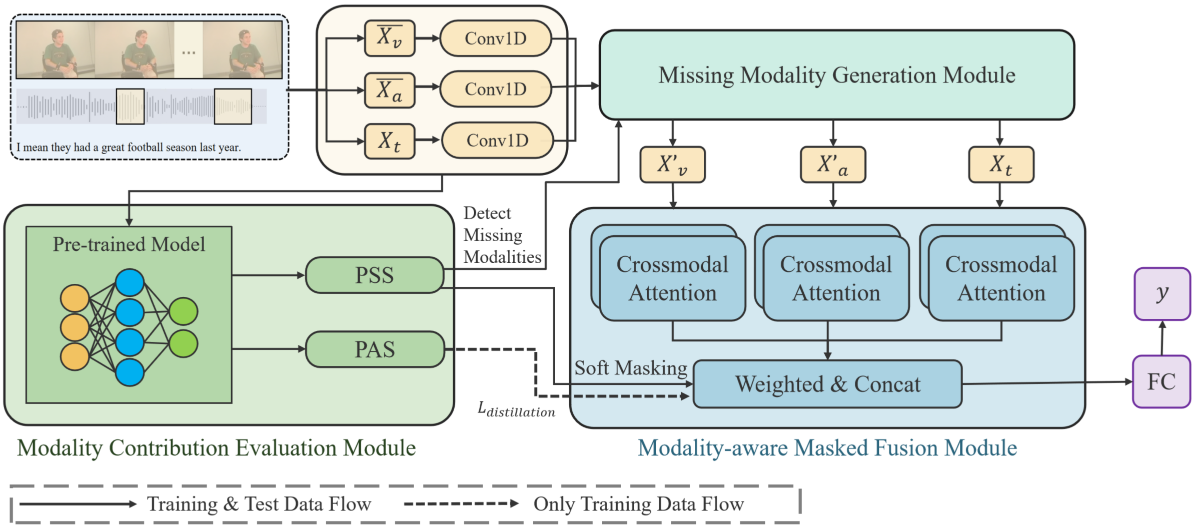

Figure 2: The overall architecture of our framework, DREAM. Conv1D refers to the 1D Convolution layer. FC refers to fully-connected layer.

contribution of each modality. Shapley value is a cooperative game theory metric that quantifies the marginal contribution of each individual to a collaborative effort within a team. By considering all possible modality combinations, it offers a theoretically robust and fair measure of each modality's unique impact on model predictions.

Prediction Shift Score. The Prediction Shift Score quantifies the ability of a modality to alter the prediction results. For a modality x_i , the marginal contribution $m^{PSS}(S, x_i)$ for a subset of modalities S is defined as follows:

$$m^{PSS}(S, x_i) = \mathbb{I}\left(\hat{y}_{S\cup\{x_i\}} \neq \hat{y}_S\right), \qquad (1)$$

where $\hat{y}_S$ denotes the prediction when only the modalities in subset S are provided as input. $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 if the condition holds and 0 otherwise.

To compute the overall contribution of a modality, it is essential to consider its impact across all possible combinations of other modalities. Specifically, the total contribution of modality x_i to the prediction is calculated as the weighted average of all possible $m^{PSS}(S, x_i)$. In this context, the PSS of modality x_i is calculated as:

$$PSS\phi_i(\mathcal{M}, x_i) =$$
$$\sum_{S\subseteq\mathcal{M}\setminus\{i\}} \frac{|S|!(|\mathcal{M}|-|S|-1)!}{|\mathcal{M}|!} m^{PSS}(S, x_i). \qquad (2)$$

Modalities that are more likely to alter the model's prediction will be assigned higher scores. In particular, the PSS of each modality can be calculated without knowing the true label.

Prediction Accuracy Score. Although PSS reflects a modality's influence on the model's decision boundary, it has the potential to assign high scores to modalities that can mislead the model. To prevent the model from overly relying on such misleading modalities, it is essential to direct the model's focus toward modalities that contribute to correct and reliable predictions. Therefore, we introduce the Prediction Accuracy Score to measure the individual contribution of each modality in achieving a correct prediction outcome. The value function of this score is adapted from (Wei et al., 2024), which is defined as:

$$V_{PAS}(S) = \begin{cases} |S| & \text{if } \hat{y_s} = y, \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

If the prediction is correct, the total payoff of a given modality subset is the number of modalities it contains. This value function assigns higher scores to the modalities that are more likely to lead the model to make the correct prediction.

The marginal contribution $m^{PAS}(S, x_i)$ of modality x_i to a subset S for computing PAS is defined as:

$$m^{PAS}(S, x_i) = V_{PAS}(S \cup x_i) - V_{PAS}(S). \qquad (4)$$

The PAS of a modality x_i is computed in the same way as PSS (see Eq. 2), as the weighted average of all possible $m^{PAS}(S, x_i)$.

3.4 Modality Enhancement Module

Instead of using manually annotated labels, the Modality Enhancement Module identifies the missing and low-contribution modalities by PSS.

Based on the computation of PSS, if a modality's PSS is zero, it implies that the introduction of the modality does not change the model's prediction. This outcome can be attributed to several underlying factors: (1) the modality is entirely missing in the given sample; (2) although present, it lacks discriminative information or the model fails to extract meaningful features from it; (3) the modality provides redundant information that is highly correlated with other modalities, offering little additional value for the prediction. Regardless of the specific cause, a modality with zero PSS reflects that it has minimal or no contribution to the model's decision, making PSS a reliable signal for identifying targets for enhancement.

The Modality Enhancement Module builds upon the idea of cross-modal reconstruction. Different from previous works (Li et al., 2024a; Wang et al., 2020b), we adopt a simple architecture to obtain a robust representation while mitigating overfitting. The specific enhancement procedure is described as follows.

Let $x_m^{(t)}$ denote the target modality to be enhanced or reconstructed. And let $\{x_1^{(a)}, x_2^{(a)}, \ldots, x_{m-1}^{(a)}\}$ represent the remaining available $M-1$ modalities. We first concatenate the representations of all modalities to construct a global multimodal context representation:

$$Z_g = [x_m^{(t)}, x_1^{(a)}, x_2^{(a)}, \ldots, x_{M-1}^{(a)}]. \tag{5}$$

Next, we generate $M-1$ partial representations, each by concatenating the target modality with one of the available modalities:

$$Z_k = [x_m^{(t)}, x_k^{(a)}], \quad \text{for } k = 1, 2, \ldots, M-1. \tag{6}$$

To model the interactions between the target modality and the multimodal context, we apply cross-modal attention (CA) modules. The target modality $x_m^{(t)}$ serves as the query, while the global and partial representations serve as the key and value inputs. We compute one global interaction:

$$I_g = CA(x_m^{(t)}, Z_g, Z_g), \tag{7}$$

and $M-1$ partial interactions:

$$I_k = CA(x_m^{(t)}, Z_k, Z_k), \quad \text{for } k = 1, 2, \ldots, M-1. \tag{8}$$

The enhanced representation of the target modality is then obtained by aggregating all interaction outputs with the original representation. The result is then normalized using Layer Normalization (LN):

$$x'_m = LN\left(x_m^{(t)} + I_g + \sum_{k=1}^{M-1} I_k\right). \tag{9}$$

This enhancement process is repeated D times, where D is a hyperparameter determined empirically.

3.5 Modality-aware Masked Fusion Module

The Modality-Aware Masked Fusion Module adopts the principle of capitalizing on strengths and suppressing weaknesses. It adaptively integrates all the original and reconstructed modality features by leveraging the PSS-based soft masking strategy and modality contribution distillation.

In the module, each modality is first processed through cross-modal attention mechanism to derive modality-aware representations $\{h_i\}_{i=1}^{M}$, serving as preliminary fusion. These representations are then concatenated and passed through a linear projection layer to predict a preliminary weight matrix w.

PSS-based Soft Masking Strategy. Generative models inevitably face the challenge of ensuring the reliability of the generated features (Zeng et al., 2022b; Guo et al., 2024). To address this problem, the influence of generated modalities should be down-weighted during fusion. PSS serves as a natural proxy, being high for original modalities and zero for reconstructed ones. We therefore introduce a PSS-based soft-masking strategy. However, directly applying PSS as a mask would exclude the generated modalities entirely from the fusion process. To prevent the loss of potentially useful information, we add a constant offset of 1 to all PSS. These adjusted PSS are then used to reweight the preliminary fusion weights, thereby modulating the influence of each modality on the final decision. The final fusion weights w' are computed as follows:

$$w' = \text{softmax}(w \times (PSS + 1)). \tag{10}$$

The fused multimodal representation H is obtained by a weighted sum of the preliminary fused representations:

$$H = \sum_{i=1}^{M} w'_i \cdot h_i. \tag{11}$$

Eventually, H is passed through a fully connected layer to produce the final prediction y.

Modality Contribution Distillation. To encourage the fusion module to rely more on the modalities that contribute positively to prediction accuracy, we introduce a distillation loss $\mathcal{L}_{distillation}$. This loss encourages the weights w to align with the corresponding PAS. Specifically, we compute this alignment using the Kullback-Leibler (KL) divergence. The loss is defined as:

$$\mathcal{L}_{distillation} = KLloss(w, PAS), \qquad (12)$$

where $KLloss(\cdot)$ refers to the KL divergence method for computing the distillation loss.

The overall training objective is to minimize the total loss $\mathcal{L}_{total}$, which is computed as:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \mathcal{L}_{distillation}, \qquad (13)$$

where $\mathcal{L}_{task}$ denotes the primary task-specific loss, which is the cross-entropy loss for classification tasks and the L1 loss for regression tasks. λ is the weight for distillation's loss.

4 Experiments

In this section, we test DREAM's performance to address the following research questions.

- Q1: How robust is DREAM to modality missingness, including both inter-modality and intra-modality missing scenarios?
- Q2: Compared to previous works, does DREAM demonstrate improved capability in addressing modality imbalance?
- Q3: Do the modules of DREAM contribute to overall performance improvements?
- Q4: To what extent does the pre-trained model influence DREAM's performance?
- Q5: What is the computational overhead of the Dynamic Modality Contribution Evaluation Module?

4.1 Experimental Setup

Datasets. We perform comprehensive evaluations on three widely-used multimodal sentiment analysis (MSA) datasets with word-level alignment: MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and IEMOCAP (Busso et al., 2008). A detailed description of datasets can be found in the Appendix A.1.

Baselines. To comprehensively evaluate the effectiveness of DREAM, we conduct extensive comparisons with a diverse set of baseline and state-of-the-art (SOTA) models. The baseline model is complete-modality method: *CubeMLP* (Sun et al.). In addition, we compare DREAM against SOTA methods specifically designed for scenarios with missing modalities, including joint learning frameworks: *CorrKD*(Li et al., 2024b) and *TransM* (Wang et al., 2020b), generative approaches: *SMIL* (Ma et al., 2021) and *GCNet* (Lian et al., 2023), as well as prompt methods: *MLPMM* (Guo et al., 2024) and *MSPs* (Jang et al., 2024).

Implementation Details. Our framework is built on the Pytorch (Paszke et al., 2017) toolbox with NVIDIA GeForce RTX 3090 GPU. The Adam optimizer (Kingma and Ba, 2015) is employed for optimization. For the pre-trained model, we adopt the backbone model of *CorrKD* (Li et al., 2024b). For detailed information on hyperparameter configurations, data preprocessing, and baseline re-implementations, please refer to Appendix A.2.

4.2 Overall Result

Q1: How robust is DREAM to modality missingness? To assess the robustness of the proposed framework, we design two sets of experiments. One simulates inter-modality missingness by removing one or more modalities during inference. The other simulates intra-modality missingness by randomly masking different proportions of frames within each modality's input sequence during both training and testing phases.

Robustness to inter-modal missingness: Table 1 presents the performance of different models under inter-modal missingness. The contents in parentheses represent the available modalities. “Avg." indicates the average performance across six missing-modality testing conditions. Our proposed framework consistently outperforms baseline and SOTA models under inter-modal missingness. On IEMOCAP dataset, when the text modality is missing, DREAM achieves an F1 score that is 6.51% higher than that of MSPs. Furthermore, DREAM achieves the highest average performance across all missing-modality scenarios. On IEMOCAP dataset, the average F1 score of DREAM is improved by 7.51% compared to that of CorrKD. This finding suggests that DREAM outperforms other models and demonstrates strong robustness to inter-modal missingness.

Dataset	Models	$\{l\}$	$\{a\}$	$\{v\}$	$\{l,a\}$	$\{l,v\}$	$\{a,v\}$	Avg.	$\{l,a,v\}$
MOSI	CubeMLP	70.87	39.85	39.32	72.15	71.95	42.31	56.03	82.57
	CorrKD	78.45	64.27	58.67	80.74	79.91	70.28	72.05	81.11
	TransM	75.02	64.54	64.48	78.99	76.56	67.24	71.13	80.87
	SMIL	77.92	**66.73**	65.32	79.01	78.06	70.34	72.89	81.26
	GCNet	79.86	63.34	65.76	77.50	80.11	69.31	72.64	81.58
	MLPMM	80.12	63.65	63.74	81.09	**81.19**	65.41	72.57	82.39
	MSPs	79.12	65.43	66.28	79.83	78.32	68.61	72.91	81.39
	DREAM (Ours)	**80.22**	59.61	**69.49**	**82.80**	79.19	**71.49**	**73.79**	**82.88**
MOSEI	CubeMLP	75.18	37.16	38.42	75.19	75.20	41.37	57.09	81.16
	CorrKD	79.12	64.75	61.03	80.08	79.63	70.46	72.51	80.49
	TransM	77.61	54.45	58.20	78.83	72.93	62.24	67.37	81.48
	SMIL	79.33	56.39	60.08	78.38	77.73	62.27	69.03	80.16
	GCNet	78.90	61.89	65.33	**80.75**	80.28	69.45	72.76	80.60
	MLPMM	79.71	68.71	69.40	80.43	80.13	69.91	74.68	81.37
	MSPs	79.91	70.23	56.19	80.16	79.98	71.04	72.91	81.26
	DREAM (Ours)	**80.03**	**77.20**	**75.24**	80.62	**81.80**	**74.80**	**78.21**	**82.11**
IEMOCAP	CubeMLP	70.07	53.06	50.22	72.10	74.25	54.06	62.29	83.37
	CorrKD	80.32	61.68	58.01	81.13	80.23	66.03	71.23	82.91
	TransM	75.30	58.35	56.21	78.07	76.27	59.88	67.34	81.37
	SMIL	79.34	59.68	56.14	81.59	**82.12**	60.28	69.85	82.21
	GCNet	80.46	61.06	59.01	81.32	79.56	61.12	70.42	83.55
	MLPMM	69.28	59.71	56.98	75.44	74.51	67.37	67.22	77.12
	MSPs	70.36	65.47	62.39	76.49	74.31	70.57	69.93	83.48
	DREAM (Ours)	**82.65**	**76.51**	**73.08**	**82.65**	80.49	**77.08**	**78.74**	**85.84**

Table 1: Performance comparison under inter-modal missingness. The evaluation metric is F1 score. The contents in parentheses represent the available modalities. The notation "{l}" indicates that only the language modality is available, while audio and visual modalities are missing. "{l, a, v}" represents the complete-modality testing condition where all modalities are available. "Avg." indicates the average performance across six missing-modality testing conditions. For the **highest** we mark in bold and the second highest we underline.

Robustness to intra-modal missingness: Figure 3 presents the performance of all models under varying levels of intra-modal missingness. All models exhibit a performance decline as the missing ratio increases, indicating sensitivity to partial modal corruption. Notably, our proposed framework demonstrates the slowest degradation trend among all the methods. Specifically, when the missing ratio exceeds 0.5, our framework consistently achieves the highest performance, underscoring its superior resilience to severe intra-modal missingness.

Q2: Does DREAM demonstrate improved capability in addressing modality imbalance? The robustness of a model to modality imbalance is reflected in its ability to effectively learn from each modality. In Table 1, the results obtained using only the text, audio, or visual modality as input reflect the model's learning proficiency for each modality. These results reveal the following findings.

All models exhibit substantially better performance when only the language modality is available, compared to when solely the visual or acoustic modalities are used. This observation highlights a clear imbalance in the predictive contribution of different modalities. Notably, our proposed model demonstrates strong performance even under scenarios where only low-contribution modalities are available. On the IEMOCAP dataset, when only the acoustic modality is retained, our model outperforms the best SOTA model *MSPs* by 11.04%.

Q3: Do the modules of DREAM contribute to overall performance improvements? To evaluate the effectiveness of the key modules in our proposed framework DREAM, we conduct ablation studies on the IEMOCAP dataset. The variant

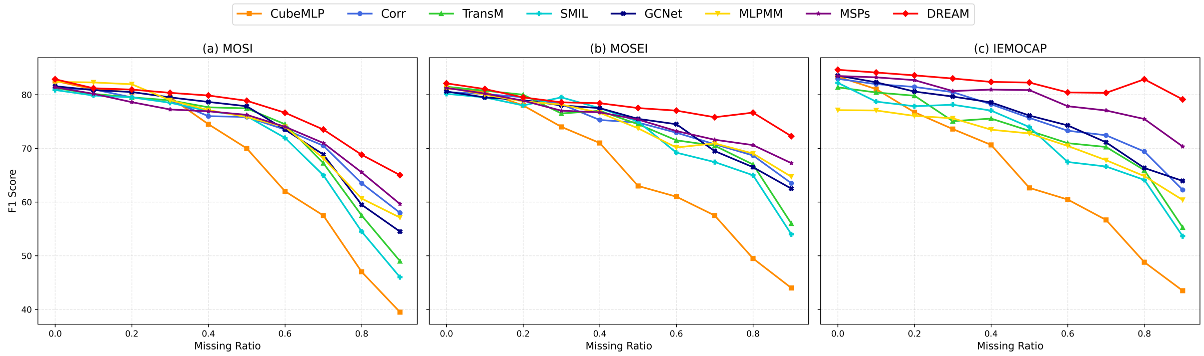

Figure 3: Comparison results of intra-modal missingness on (a) MOSI, (b) MOSEI, and (c) IEMOCAP at various missing ratios.

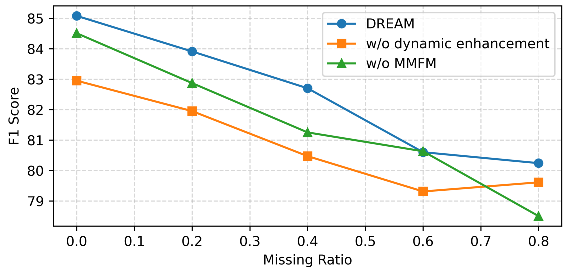

Figure 4: Results of ablation experiments under multiple modalities missing situation.

w/o dynamic enhancement removes the constraint for low contribution modalities and instead reconstructs all modalities indiscriminately. The variant **w/o MMFM** removes the adaptive fusion process and instead fuses multiple modalities based on predicted weights. In our experiments, we simulate both intra- and inter-modality missingness at the same time. Further details of the experimental settings are provided in the Appendix A.4.

The results are presented in Figure 4. As shown in the results, the complete DREAM model consistently outperforms both ablated variants across all missing rates. This confirms that both the modality enhancement module and modality-aware masked fusion module play important roles in improving robustness under partial or severe missing conditions.

Q4: To what extent does the pre-trained model influence DREAM's performance? We further conduct ablation studies on the pre-trained model used within the DREAM framework to investigate the extent to which DREAM relies on the quality of the pre-trained backbone. Specifically, we compare three alternative models: (1) Self-MM (Yu et al., 2022), a strong performer on fully observed multimodal datasets; (2) CorrKD, which shares the same architecture as our main setup but is insufficiently trained; and (3) a simple multilayer perceptron (MLP) baseline.

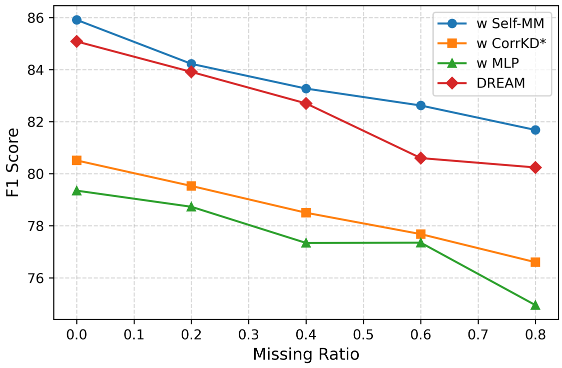

Figure 5: Results of ablation experiments with different pre-trained models under multiple modalities missing situation.

As shown in Figure 5, DREAM consistently benefits from stronger pre-trained models, with Self-MM delivering the best overall performance. Yet even with weaker backbones such as MLP, DREAM remains competitive. These results suggest that while DREAM leverages high-quality pre-trained models for enhanced accuracy, it retains robustness with lightweight or suboptimally trained alternatives. Therefore, depending on specific deployment needs, practitioners may flexibly trade off between accuracy and efficiency by choosing pre-trained models of varying complexity. This finding highlights the robustness and adaptability of DREAM across different model scales and capacities.

Q5: What is the computational overhead of the Dynamic Modality Contribution Evalua-

tion Module? The Dynamic Modality Contribution Evaluation Module forms the cornerstone of the DREAM framework, as it precisely quantifies the contribution of each modality based on a pre-trained model. By definition, computing the contributions of $\mathcal{M}$ modalities requires 2^M predictions, introducing potential computational expense. To assess the cost, we compared the overall runtime of our framework against several competitive baselines, including CubeMLP, TransM, and GCNet. We also analyzed the cost introduced by the dynamic evaluation. The results are summarized in Table 2.

As shown in Table 2, for scenarios with three modalities, our framework maintains acceptable—even efficient—training and testing times. This efficiency largely stems from our choice of relatively lightweight pre-trained models, which substantially reduce computational overhead. Moreover, the dynamic evaluation strategy alleviates the need for additional manual annotations, further lowering the cost.

Nevertheless, for larger modality sets ($\mathcal{M} > 4$), exact calculation becomes computationally infeasible. In such cases, approximation techniques like the Monte-Carlo method (Luo et al., 2024) or the Shapley Additive Explanations method (Singh and Chaturvedi, 2024) can be integrated. Our findings in Section 4.2 (Q4), which show that DREAM performs well even with less accurate contribution estimates, support the viability of using such approximations to ensure scalability.

	IEMOCAP		MOSI		MOSEI	
Models	train	test	train	test	train	test
DREAM(ours)	132.16	1.64	83.63	1.22	793.41	3.74
pre-trained	16.23	1.21	6.25	0.93	85.47	2.13
CubeMLP	121.95	1.13	81.10	1.05	764.61	2.57
TransM	773.53	1.07	389.03	1.16	4106.38	2.95
GCNet	207.42	1.09	124.32	0.95	1476.18	3.41

Table 2: Runtime comparison across IEMOCAP, MOSI, and MOSEI datasets. The results of DREAM (ours) represent the overall running time of the framework. "Pre-trained" refers to the total computation time of the pre-trained model.

5 Conclusion

In this paper, we present DREAM, a framework that addresses both missing and imbalanced modalities. DREAM dynamically identifies modalities with low contributions and reconstructs them to enhance overall representation quality. DREAM adaptively fuses both the original and reconstructed modalities based on their estimated contributions. Importantly, the entire learning process is carried out without the need for full-modality supervision or explicit annotations indicating missing modalities. Experiments on three benchmark datasets demonstrate that DREAM achieves strong performance and robustness with missing and imbalanced modalities. The results of the ablation studies further highlight the effectiveness of DREAM's dynamic modality contribution recognition and its robustness to different pre-trained models.

Limitations

Although our proposed method demonstrates strong robustness under both inter-modal and intra-modal missingness scenarios, there remain limitations to be addressed in future work. In particular, our approach does not leverage recent advances in large language models (LLMs), which have shown remarkable capabilities in cross-modal reasoning and representation learning. Incorporating LLMs such as GPT or BERT-derived architectures could further enhance the semantic understanding of the textual modality, and potentially benefit multi-modal fusion through pre-trained multimodal representations. Exploring this integration remains an important direction for improving both accuracy and generalization.

Acknowledgments

This work is supported by National Natural Science Foundation of China NSFC (No.92370201, No. 62272165) and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Pierre Baldi. 2012. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 37–49, Bellevue, Washington, USA. PMLR.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S.

Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, page 335–359.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 8632–8656. PMLR.

Christian Ganhör, Marta Moscati, Anna Hausberger, Shah Nawaz, and Markus Schedl. 2024. A multi-modal single-branch embedding network for recommendation in cold-start and missing modality scenarios. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, pages 380–390. ACM.

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

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9180–9192. Association for Computational Linguistics.

Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. 2020. *Knowledge Distillation from Multi-modal to Mono-modal Segmentation Networks*, page 772–781.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9226–9259. PMLR.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. 2024. Towards robust multimodal prompting with missing modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 8070–8074. IEEE.

Donggeun Kim and Taesup Kim. 2024. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVI*, volume 15144 of *Lecture Notes in Computer Science*, pages 171–187. Springer.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14943–14952. IEEE.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. 2024a. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 10074–10082. AAAI Press.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024b. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 12458–12468. IEEE.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6631–6640. IEEE.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8419–8432.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.*, 56(10):264.

Muhammad Irzam Liaqat, Shah Nawaz, Muhammad Zaigham Zaheer, Muhammad Saad Saeed, Hassan Sajjad, Tom De Schepper, Karthik Nandakumar, Muhammad Haris Khan, Ignazio Gallo, and Markus Schedl. 2025. Chameleon: A multimodal learning framework robust to missing modalities. *Int. J. Multim. Inf. Retr.*, 14(2):21.

Wen Luo, Yu Xia, Tianshu Shen, and Sujian Li. 2024. Shapley value-based contrastive alignment for multimodal information extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 5270–5279. ACM.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18156–18165. IEEE.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. SMIL: multimodal learning with severely missing modality. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2302–2310. AAAI Press.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8228–8237. IEEE.

Shashank Kumar Singh and Amrita Chaturvedi. 2024. An efficient multi-modal sensors feature fusion approach for handwritten characters recognition using shapley values and deep autoencoder. *Eng. Appl. Artif. Intell.*, 138:109225.

Hao Sun, Yen-Wei Chen, Hongyi Wang, Jiaqing Liu, and Lanfen Lin. Cubemlp: A mlp-based model for multimodal sentiment analysis and depression estimation.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: A mlp-based model for multimodal sentiment analysis and depression estimation. *CoRR*, abs/2207.14087.

Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. 2024. Redcore: relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4971–4980. IEEE Computer Society.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics.

Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part IV*, volume 14223 of *Lecture Notes in Computer Science*, pages 216–226. Springer.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020a. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12692–12702. Computer Vision Foundation / IEEE.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020b. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2514–2520. ACM / IW3C2.

Robert James Weber. 1988. *Probabilistic values for games*, page 101–120. Cambridge University Press.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. 2024. Enhancing multimodal cooperation via sample-level modality valuation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 27328–27337. IEEE.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 24043–24055. PMLR.

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. 2024. Drfuse: learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and*

Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2022. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 10790–10797.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv: Computation and Language,arXiv: Computation and Language*.

Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022a. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1545–1554. ACM.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. 2022b. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2924–2934. Association for Computational Linguistics.

Changqing Zhang, Zongbo Han, Yajie Cui, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. 2019. Cpm-nets: Cross partial multi-view networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 557–567.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024a. Multimodal representation learning by alternating unimodal adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 27446–27456. IEEE.

Zheyu Zhang, Gang Yang, Yueyi Zhang, Huanjing Yue, Aiping Liu, Yunwei Ou, Jian Gong, and Xiaoyan Sun. 2024b. Tmformer: Token merging transformer for brain tumor segmentation with missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7414–7422.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2608–2618. Association for Computational Linguistics.

Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu. 2024. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7534–7542.

A Appendix

A.1 Dataset

This study utilizes the IEMOCAP (Busso et al., 2008), CMU-MOSI (Zadeh et al., 2016), and CMU-MOSEI (Bagher Zadeh et al., 2018) datasets.

- **The IEMOCAP dataset** is available for non-commercial research purposes under a custom license from the USC SAIL Lab.
- **The CMU-MOSI dataset** is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.
- **The CMU-MOSEI dataset** is provided for academic research purposes under a custom license via the CMU Multimodal SDK.

All datasets were accessed and used in compliance with their respective licensing agreements.

The IEMOCAP dataset includes 4,453 video samples, with predefined splits of 2,717 for training, 798 for validation, and 938 for testing. It focuses on four categorical emotions: happy, sad, angry, and neutral. The MOSI dataset comprises 2,199 short monologue video segments, partitioned into 1,284 for training, 229 for validation, and 686 for testing. MOSEI, a larger-scale corpus, contains 22,856 video clips divided into 16,326 training, 1,871 validation, and 4,659 testing samples. Both datasets are annotated with sentiment scores ranging from -3 (strongly negative) to +3 (strongly positive). For performance evaluation on MOSI and MOSEI, we adopt the weighted F1 score calculated based on binary (positive/negative) sentiment classification. The IEMOCAP dataset includes 4,453 video samples, with predefined splits of 2,717 for training, 798 for validation, and 938 for testing. It focuses on four categorical emotions: happy, sad, angry, and neutral.

A.2 Hyperparameter and Baseline Implementation

For MOSI, MOSEI, and IEMOCAP, the detailed hyperparameter settings are as follows: the learning rates are 1e-4, 1e-4, 2e-4. For all three datasets, the batch sizes are 32, the epoch numbers are 30, and the embedding dimension is 200. In modality enhancement module, the reconstruction process is repeated 2 times. Hyperparameters are chosen based on the validation set performance. For samples with missing modalities, the corresponding feature vectors are replaced with zeros. To ensure a fair comparison, all state-of-the-art baselines are re-implemented using publicly released codebases. The final results are obtained by averaging the outcomes of five runs with different random seeds.

A.3 Parameter Stastics

To better understand the computational complexity of our proposed framework, we report the number of trainable parameters used in our experiments across different datasets. Table 3 summarizes the parameter counts for the pre-trained model, the DREAM framework, and the total combined parameters. It is worth noting that while DREAM introduces additional parameters beyond the pre-trained model, the total model size remains within a reasonable range for practical applications.

Component	**MOSI**	**MOSEI**	**IEMOCAP**
Pre-trained model	77,762	93,890	95,472
DREAM	3,054,116	3,070,916	3,072,323
Total	**3,131,878**	**3,164,806**	**3,167,795**

Table 3: Number of trainable parameters on different datasets.

A.4 Experimental Setup for Heterogeneous Modality Missingness

As detailed in our ablation experiments (Section 4.2, Q3 and Q4), we explored a more challenging scenario where both inter-modal and intra-modal missingness are present simultaneously. Specifically, when the overall missing rate is set to $\eta\%$, both the training and testing sets contain $\eta\%$ of samples with missing modalities, while the remaining $1-\eta\%$ contain all the modalities. Among all the samples:

- $\frac{2\eta}{5}\%$ suffer from inter-modal missingness,
- $\frac{2\eta}{5}\%$ suffer from intra-modal missingness,
- $\frac{\eta}{5}\%$ suffer from both types of missingness concurrently,
- $1-\eta\%$ contain all the modalities.

This distribution better simulates real-world multimodal learning scenarios, where missingness is heterogeneous and unpredictable.