Supervised Attention Mechanism for Low-quality Multimodal Data

Sijie Mai[1] **and Shiqin Han**[1] **and Haifeng Hu**[2*]
[1] School of Computer Science, South China Normal University
[2] School of Electronics and Information Technology, Sun Yat-sen University
{sijiemai,20222121019}@m.scnu.edu.cn, huhaif@mail.sysu.edu.cn

Abstract

In practical applications, multimodal data are often of low quality, with noisy modalities and missing modalities being typical forms that severely hinder model performance, robustness, and applicability. However, current studies address these issues separately. To this end, we propose a framework for multimodal affective computing that jointly addresses missing and noisy modalities to enhance model robustness in low-quality data scenarios. Specifically, we view missing modality as a special case of noisy modality, and propose a supervised attention framework. In contrast to traditional attention mechanisms that rely on main task loss to update the parameters, we design supervisory signals for the learning of attention weights, ensuring that attention mechanisms can focus on discriminative information and suppress noisy information. We further propose a ranking-based optimization strategy to compare the relative importance of different interactions by adding a ranking constraint for attention weights, avoiding training noise caused by inaccurate absolute labels. The proposed model consistently outperforms state-of-the-art baselines on multiple datasets under the settings of complete modalities, missing modalities, and noisy modalities.

1 Introduction

In the real world, objects and events are often depicted through multiple modalities, and humans perceive the world through various senses (seeing, hearing, etc), which underscores the importance of integrating data from different sources (Baltrušaitis et al., 2019). Consequently, multimodal affective computing (MAC) (Poria et al., 2017), aiming to effectively integrate language, acoustic, and visual cues from speakers to comprehensively understand and predict human sentiments, viewpoints, mental states, intentions, etc., is a promising area in multimodal learning with great application potential.

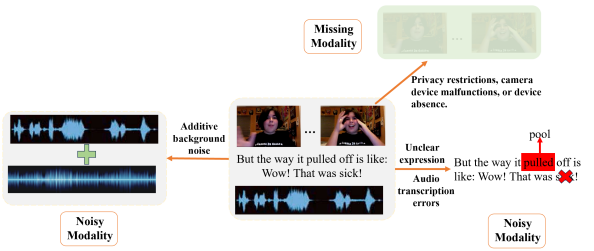


Figure 1: An example of missing and noisy modalities.

However, in practical applications, multimodal data can be of low quality, with noisy modality and missing modality data being typical forms that significantly hinder the performance, robustness and applicability of the model (Zhang et al., 2024; Liu et al., 2024). As shown in Figure 1, missing modality is often caused by unavailable data collection equipment or sensor failures, while modality noise can result from background interference, sensor noise, or data transmission errors. Current studies often address missing and noisy modality problems separately. However, since both problems are common and often occur simultaneously, separate handling limits the application scope and robustness of the model.

Consequently, we propose a multimodal framework to jointly address missing and noisy modalities, aiming to enhance the robustness of the model in low-quality data scenarios. We view missing modality as a special type of noisy modality, where the noise generation pattern and affected modality are known. Then we propose a supervised attention mechanism for low-quality multimodal learning (SAM-LML). Attention mechanisms are widely-used for multimodal fusion and are the basic units for large language models (Vaswani et al., 2017; Touvron et al., 2023; Radford et al., 2021; Chay-intr et al., 2025). However, conventional attention mechanisms lack explicit supervisory signals where the update of parameters depends solely on the loss of main task. This might lead to the

*Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21366–21386
November 4-9, 2025

learning of noisy correlations or the highlighting of non-discriminative information (see Section 4.6), potentially leading to poorer interpretability and suboptimal performance. In contrast, SAM-LML explicitly designs supervisory signals for the learning of attention weights, enabling more accurate capturing of important intra- and inter-modal interactions. This effectively suppresses modality noise and enhances the interpretability of attention mechanisms. Additionally, the ability of attention mechanism to handle variable-length sequences makes it well-suited for addressing missing modalities.

Particularly, SAM-LML involves a supervised intra-modal attention mechanism, which is suitable for handling partial feature missing or noise within unimodal sequences. By incorporating noise addition, modality decomposition replacement, attention matrix variance and context constraints, it designs supervisory signals for the learning attention weights, helping to suppress irrelevant information and focus on important intra-modal interactions. Furthermore, the supervised inter-modal attention applies variance constraints, noise addition and modality mixing operations to learn discriminative inter-modal interactions and suppress low-quality modalities, which is designed for scenarios with complete missing or noisy modalities.

However, directly defining absolute labels for attention weights is often inappropriate due to the difficulty of the definition of absolute labels, which can leads to training noise. To address this, we further propose a ranking-based optimization strategy to guide the model to compare the relative importance between different interactions. It seeks to leverage the flexibility and expressive power of deep learning models to automatically learn attention weights, and only requires the learned weights to follow a constraint introduced by the ranking relationships of weights, avoiding issues with inaccurate absolute labels. Moreover, the ranking-based training strategy explicitly compares the importance of interactions, aligning with the essence of attention mechanisms, i.e., focusing on more informative and important interactions.

Our main contributions are listed as below:

- We innovatively address noisy modality and missing modality problems in a unified framework, which enhances the model's potential for application in real-world scenarios.
- We consider missing modality as a special type of noisy modality, and design supervisory signals for the learning of attention weights to capture important intra- and inter-modal interactions more accurately. This effectively suppresses modality noise and boosts the interpretability of attention mechanisms.
- We propose a ranking-based optimization strategy to compare the relative importance between different interactions, which only requires the learned weights to follow a ranking constraint and can avoid training noise caused by inaccurate absolute labels.
- SAM-LML outperforms state-of-the-art methods on multiple MAC datasets under the settings of complete modalities, missing modalities, and noisy modalities. The visualizations show that SAM-LML can enhance the interpretability of attention mechanisms.

2 Related Work

2.1 Missing Modality

The methods for addressing missing modality issue can be roughly categorized into four categories: (1) **Data augmentation methods** simulate the absence of modality during training to improve generalizability via introducing modality drop and noise (Lin and Hu, 2024; Hazarika et al., 2022). However, simple data augmentation methods often lead to a decrease in the performance of complete multimodal fusion; (2) **Alignment-based methods** align the representations of incomplete and complete modalities through contrastive learning (Lin and Hu, 2023; Poklukar et al., 2022), canonical correlation analysis (Andrew et al., 2013; Sun et al., 2020), or knowledge distillation (Li et al., 2024), etc. However, these methods require training samples to have complete modalities and often need to additionally train a teacher network; (3) **Reconstruction-based methods** focus on generating missing modality features using generative models, including autoencoders (Tran et al., 2017; Zeng et al., 2022), variational autoencoders (Shi et al., 2019), graph completion networks (Lian et al., 2023), diffusion models (Wang et al., 2023), and prompt-tuning strategies (Guo et al., 2024). These methods effectively compensate for missing modality information, but they often suffer from high complexity; (4) **Architecture-based methods** leverage architectures that can naturally handle an arbitrary number of modalities (such as attention mechanisms, mixture-of-expert models, ensemble

algorithms) to address missing modalities (Xu et al., 2024a; Deng et al., 2025; Xue and Marculescu, 2023). For example, Qian and Wang (2023) design a masked attention mechanism that replaces missing values with zeros or negative infinity. However, they do not design specific strategies to enhance model performance in missing modality scenarios.

Compared to these methods, SAM-LML views missing modality as a special type of noisy modality and designs supervisory signals for the learning of attention weights, enhancing the model's adaptability and robustness to missing modalities.

2.2 Noisy Modality

The key methods for handling modality noise mainly fall into three categories: (1) **Noise augmentation methods** simulate real-world noise by adding noise to features or raw inputs and designing defense mechanisms accordingly (Hazarika et al., 2022; Mao et al., 2023). However, designing defense mechanisms for each type of noise can be complex and has limited generalizability; (2) **Representation regularization methods** use techniques such as tensor rank minimization to extract discriminative representations from noisy features (Liang et al., 2019). For example, researchers apply the information bottleneck principle to filter modality noise by minimizing mutual information between encoded features and inputs (Mai et al., 2023c; Federici et al., 2020). However, they rely on assumptions that may not hold in real-world scenarios; (3) **Noise identification and filtering methods** aim to identify and suppress noisy information (Gong et al., 2025). For example, Xue et al. (2023) propose a multi-level attention map network to filter modality noise before fusion. However, attention-based methods lack explicit supervisory signals to learn accurate modality weights, making them prone to focusing on noisy information. Multimodal Boosting (Mai et al., 2024) uses unimodal predictive losses as supervisory signals to calculate absolute labels for modality contributions, thus identifying noisy modalities. However, absolute labels are difficult to define accurately, which may introduce training noise. In contrast, we propose a ranking-based training strategy to compare the relative importance of interactions, which avoids inaccurate learning and aligns with the nature of attention: focusing on more important information.

Moreover, research that can simultaneously address noisy and missing modalities remains underdeveloped. Since missing and noisy modalities often coexist in real-world low-quality data scenarios, we aim to simultaneously address these two issues to expand the applicability of the model.

3 Algorithm

The diagram of SAM-LML is shown in Figure 2. SAM-LML is evaluated on multiple tasks of MAC, including multimodal sentiment analysis (MSA) (Zadeh et al., 2016), multimodal humor detection (MHD) (Hasan et al., 2019), and multimodal sarcasm detection (MSD) (Castro et al., 2019). The input is a video segment of a speaker described by acoustic (a), visual (v), and language (l) modalities. The input feature sequences are denoted as $\{\boldsymbol{Z}_m \in \mathbb{R}^{T\times d_m} | m \in \{a, v, l\}\}$, where T is the sequence length and d_m is the feature dimensionality. When one modality is absent, we use Gaussian noise to serve as the features of that modality.

3.1 Intra-Modal Attention

Given multiple unimodal features $\boldsymbol{U}_m \in \mathbb{R}^{T\times d}$ output by unimodal networks (see Section A.1), we first apply regular self-attention modules (Vaswani et al., 2017) to explore intra-modal interactions:

$$\boldsymbol{A}_m = \frac{\boldsymbol{Q}\boldsymbol{K}^R}{\sqrt{d}} \in \mathbb{R}^{T\times T} \tag{1}$$

$$\boldsymbol{X}_m = LN(Softmax(\boldsymbol{A}_m)\boldsymbol{V}) \tag{2}$$

where $\boldsymbol{Q} = \boldsymbol{U}_m\boldsymbol{W}_{m,q}$, $\boldsymbol{K} = \boldsymbol{U}_m\boldsymbol{W}_{m,k}$, $\boldsymbol{V} = \boldsymbol{U}_m\boldsymbol{W}_{m,v}$, $\boldsymbol{A}_m$ denotes the attention matrix for modality m, R is the matrix transpose operation, LN is the layer normalization, and $\boldsymbol{X}_m \in \mathbb{R}^{T\times d}$ is the high-level unimodal representation for modality m.

However, current attention mechanisms lack explicit supervisory information for learning attention weights, making them prone to focusing on noise and irrelevant interactions. Our experiments show that conventional attention mechanisms sometimes fail to detect noise (see Section 4.6 and A.7). Unlike traditional methods, SAM-LML enables supervised learning of attention weights, enhancing the interpretability of attention mechanisms and ensuring more accurate detection of noise. Specifically, we design the following supervisory signals and constraints for intra-modal attention:

(1) **Noise Addition**: This operation randomly adds noise to the features of keys. If the model can assign lower weights to noisy information, it can have a clearer understanding of informative interactions. A straightforward supervised approach is

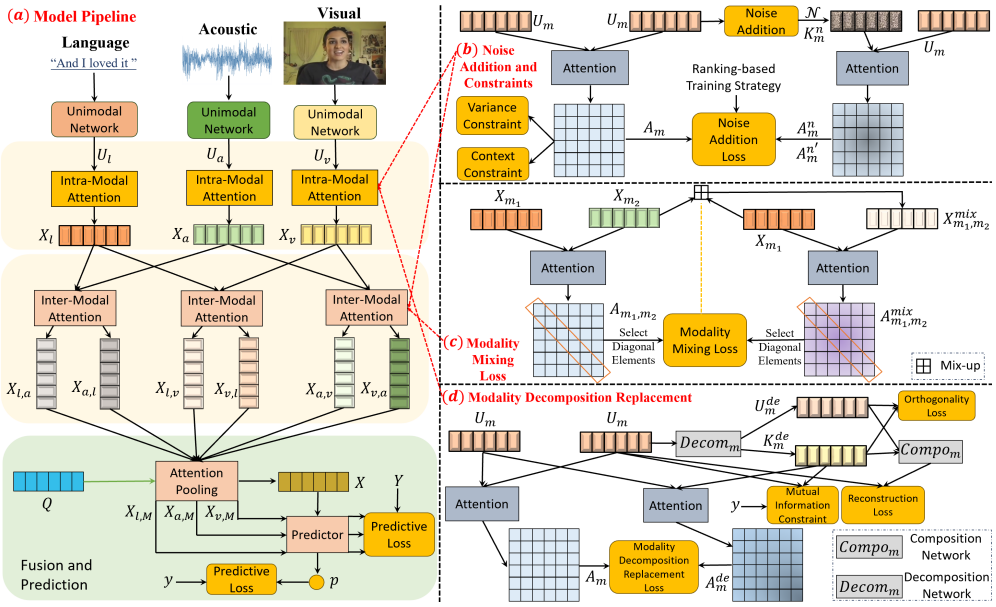


Figure 2: Diagram of the overall framework.

to encourage the attention weights between queries and noisy features as small as possible. However, this strategy does not provide supervisory signals to the weights between queries and original keys, and cannot ensure the original weights will not also be minimized. Moreover, noisy features might still preserve discriminative modality information. To address this, we propose a **ranking-based optimization strategy**, encouraging the attention values between queries and original keys to be greater than those between queries and noisy data plus a pre-defined margin. It guides attention mechanisms to compare and learn the relative importance of interactions, which aligns with the nature of attention mechanisms, i.e., focusing on more informative and important interactions. Compared to methods that assign absolute labels to attention weights (Mai et al., 2024; Li et al., 2018), our strategy can avoid the inaccuracies of absolute labels assigned based on human experience, and retains the expressiveness and flexibility of deep neural networks to automatically determine more accurate weights under ranking constraints. Specifically, the ranking-based noise addition loss $\mathcal{L}_m^n$ is defined as:

$$\boldsymbol{A}_m^n = \frac{\boldsymbol{U}_m \boldsymbol{W}_{m,q} (\mathcal{N} \boldsymbol{W}_{m,k})^R}{\sqrt{d}} \in \mathbb{R}^{T \times T} \quad (3)$$

$$\mathcal{L}_m^n = max(\boldsymbol{A}_m^n - \boldsymbol{A}_m + \gamma_1,\ 0) \quad (4)$$

where $\mathcal{N} \in \mathbb{R}^{T \times d}$ denotes the Gaussian noise, and γ_1 is the margin (a hyperparameter) that controls the magnitude of differences between weights. $\mathcal{L}_m^n$ forces the original attention weights to be larger than noisy weights plus the margin γ_1, which guides the model to explicitly identify noisy interactions and thus implicitly helps the model better understand the nature of informative interactions.

Additionally, to simulate real-world noise, we add Gaussian noise to the key features to generate augmented noisy features, and encourage the attention values between queries and augmented noisy features to be greater than those between queries and pure noisy data:

$$\boldsymbol{K}_m^{n'} = \lambda \cdot \boldsymbol{U}_m + (1 - \lambda) \cdot \mathcal{N} \quad (5)$$

$$\mathcal{L}_m^n \longleftarrow \mathcal{L}_m^n + max(\boldsymbol{A}_m^n - \boldsymbol{A}_m^{n'} + \gamma_1,\ 0) \quad (6)$$

where $\boldsymbol{A}_m^{n'}$ is the attention matrix of $\boldsymbol{U}_m$ and $\boldsymbol{K}_m^{n'}$, and λ is a random number between 0 and 1. In this way, the model can learn to extract discriminative information from augmented features (augmented features still contain modality information).

(2) **Modality Decomposition Replacement**: To further enable the model to recognize discriminative information, the proposed modality decomposition replacement orthogonally decomposes modality representations into label-irrelevant and discriminative components. The original key representations are replaced with the label-irrelevant ones, and the model is required to assign higher attention weights to the original key-query pairs than to the replaced key-query pairs:

$$\boldsymbol{A}_m^{de} = \frac{\boldsymbol{U}_m \boldsymbol{W}_{m,q} (\boldsymbol{K}_m^{de} \boldsymbol{W}_{m,k})^R}{\sqrt{d}} \in \mathbb{R}^{T \times T} \quad (7)$$

$$\mathcal{L}_m^{de} = max(\boldsymbol{A}_m^{de} - \boldsymbol{A}_m + \gamma_1,\ 0) \quad (8)$$

where $\boldsymbol{K}_m^{de}$ denotes the label-irrelevant representation. By minimizing $\mathcal{L}_m^{de}$, we encourage the model to explicitly identify more discriminative information and interactions. The proposed modality decomposition operation is defined as:

$$\begin{aligned} \boldsymbol{K}_m^{de}, \boldsymbol{U}_m^{de} &= Decom_m(\boldsymbol{U}_m;\ \theta_{decom_m}) \\ \mathcal{L}_m^{or} &= ||\boldsymbol{U}_m^{de}(\boldsymbol{K}_m^{de})^R||_2 \end{aligned} \tag{9}$$

$$\mathcal{L}_m^{MI} = MI(y, \boldsymbol{K}_m^{de}) + \alpha \cdot max(||\boldsymbol{U}_m - \boldsymbol{K}_m^{de}||^2 - \epsilon, 0) \tag{10}$$

$$\begin{aligned} \hat{\boldsymbol{U}}_m &= Compo_m(\boldsymbol{K}_m^{de}, \boldsymbol{U}_m^{de};\ \theta_{compo_m}) \\ \mathcal{L}_m^{re} &= ||\boldsymbol{U}_m - \hat{\boldsymbol{U}}_m||_2 \end{aligned} \tag{11}$$

where $Decom_m$ and $Compo_m$ denote the decomposition and composition networks for modality m, $\boldsymbol{U}_m^{de}$ is the label-relevant representation, y is the ground true label, and $MI(y, \boldsymbol{K}_m^{de})$ quantifies the mutual information between the label-irrelevant representation $\boldsymbol{K}_m^{de}$ and y. The hyperparameter α is the weight of loss, ϵ is the maximum distance between $\boldsymbol{U}_m$ and $\boldsymbol{K}_m^{de}$ (which ensures $\boldsymbol{K}_m^{de}$ and the original representation $\boldsymbol{U}_m$ maintain a degree of similarity but are not identical), and $\hat{\boldsymbol{U}}_m$ is the reconstructed representation. The orthogonality loss $\mathcal{L}_m^{or}$ in Eq. 9 enforces orthogonality between $\boldsymbol{U}_m^{de}$ and $\boldsymbol{K}_m^{de}$. The mutual information constraint $\mathcal{L}_m^{MI}$ ensures that $\boldsymbol{K}_m^{de}$ lacks discriminative information by minimizing $MI(y, \boldsymbol{K}_m^{de})$, and maintains relevance to $\boldsymbol{U}_m^{de}$ by minimizing $max(||\boldsymbol{U}_m - \boldsymbol{K}_m^{de}||^2 - \epsilon, 0)$, thus preventing $\boldsymbol{K}_m^{de}$ from only containing information unrelated to $\boldsymbol{U}_m$. The reconstruction loss $\mathcal{L}_m^{re}$ forces $\boldsymbol{U}_m^{de}$ and $\boldsymbol{K}_m^{de}$ to fully encompass the information in $\boldsymbol{U}_m$, preventing information loss during decomposition. These constraints enable the model to derive label-irrelevant yet $\boldsymbol{U}_m$-related representation $\boldsymbol{K}_m^{de}$.

To minimize the mutual information $MI(y, \boldsymbol{K}_m^{de})$ in Eq. 10 for effective modality decomposition, we can apply the information bottleneck principle (Mai et al., 2023c; Tishby et al., 2000) which however adds complexity to the model and relies on assumptions that might not hold in real-world scenarios. Therefore, we design a simpler and effective operation that encourages the prediction derived from $\boldsymbol{K}_m^{de}$ to be the same as the prediction derived from Gaussian noise, so that $\boldsymbol{K}_m^{de}$ cannot contain discriminative information with respect to the label:

$$\begin{aligned} MI(y, \boldsymbol{K}_m^{de}) = ||&Predictor(\boldsymbol{K}_m^{de}; \theta_{pre}) - \\ &Predictor(\mathcal{N}; \theta_{pre})||^2 \end{aligned} \tag{12}$$

where $\mathcal{N}$ is the Gaussian noise, and $Predictor$ is the final predictor parameterized by θ_{pre}.

The final loss for modality decomposition replacement is:

$$\mathcal{L}_m^{de} \longleftarrow \mathcal{L}_m^{de} + \alpha_{de} \cdot (\mathcal{L}_m^{or} + \mathcal{L}_m^{MI} + \mathcal{L}_m^{re}) \tag{13}$$

where α_{de} is the weight of decomposition losses.

(3) **Variant Constraint**: To prevent the learned attention weights from being overly similar and losing the ability to distinguish different interactions, we impose a variance constraint on the attention matrix, ensuring that the variance of the elements within the attention matrix exceeds a certain threshold. In this way, we enable different attention weights to have sufficient distinguishability:

$$\mathcal{L}_m^{var} = max(0, -Std(Softmax(\boldsymbol{A}_m)) + \beta) \tag{14}$$

where Std function calculates the variance of each row in the attention matrix. The variance constraint loss ensures that the variance of each row in the attention matrix exceeds the hyperparameter β.

(4) **Context Constraint**: In the self-attention matrix, due to contextual correlations, the features of each time step should be highly correlated with the features of its adjacent time steps. In other words, elements near the diagonal of attention matrix should have larger correlation values. To achieve this, we propose a context constraint:

$$\mathcal{L}_m^{con} = max(0, mean(\boldsymbol{A}_m) + \gamma - mean_{dia_\rho}(\boldsymbol{A}_m)) \tag{15}$$

where $mean(\boldsymbol{A}_m)$ calculates the average of all elements in each row of the attention matrix, while $mean_{dia_\rho}(\boldsymbol{A}_m)$ calculates the average of the η elements near the diagonal in each row. $\mathcal{L}_m^{con}$ ensures that the average of ρ elements near the diagonal in each row is greater than the average of all elements in that row plus the margin γ.

3.2 Inter-Modal Attention

After using supervised intra-modal attention to learn unimodal representation $\boldsymbol{X}_m$, for each pair of modalities, we take one modality as the query and the other as the key and value, and employ inter-modal attention to learn cross-modal interactions, extracting meaningful modality correlations and complementary information to learn discriminative multimodal representations. For modalities m_1 and m_2, the procedure can be expressed as:

$$\boldsymbol{A}_{m_1,m_2} = \frac{\boldsymbol{X}_{m_1}\boldsymbol{W}_{m_1,q_M}(\boldsymbol{X}_{m_2}\boldsymbol{W}_{m_1,k_M})^R}{\sqrt{d}} \tag{16}$$

$$\boldsymbol{X}_{m_1,m_2}=LN(Softmax(\boldsymbol{A}_{m_1,m_2})\boldsymbol{X}_{m_2}\boldsymbol{W}_{m_1,v_M})) \tag{17}$$

where $\boldsymbol{X}_{m_1,m_2} \in \mathbb{R}^{T\times d}$ represents the cross-modal representation. To simplify calculation and reduce complexity, we conduct average pooling on cross-modal representations at the time dimension:

$$\hat{\boldsymbol{X}}_{m_1,M}=\sum_{m_2}^{m_2\neq m_1}\{\boldsymbol{X}_{m_1,m_2}\}_{\oplus} \tag{18}$$

$$\boldsymbol{X}_{m_1,M}=\frac{1}{T\cdot(|\mathcal{M}|-1)}\sum_{t=1}^{T\cdot(|\mathcal{M}|-1)}(\hat{\boldsymbol{X}}_{m_1,M})_t \tag{19}$$

where $\bigoplus$ is the concatenation operation at time dimension, $|\mathcal{M}|$ represents the number of modalities, $\boldsymbol{X}_{m_1,M} \in \mathbb{R}^{1\times d}$ denotes the concatenated cross-modal representation where modality m_1 is the query. After obtaining multiple cross-modal representations, we define a learnable query $\boldsymbol{Q} \in \mathbb{R}^{1\times T}$ for attention pooling to acquire the multimodal representation $\boldsymbol{X}$ and then generate prediction:

$$\hat{\boldsymbol{X}}=\sum_{m_1}\{\boldsymbol{X}_{m_1,M}\}_{\oplus} \in \mathbb{R}^{|\mathcal{M}|\times d} \tag{20}$$

$$\boldsymbol{A}_M=\frac{\boldsymbol{Q}\boldsymbol{W}_{M,q}(\hat{\boldsymbol{X}}\boldsymbol{W}_{M,k})^R}{\sqrt{d}} \in \mathbb{R}^{1\times|\mathcal{M}|} \tag{21}$$

$$\boldsymbol{X}=LN(Softmax(\boldsymbol{A}_M)\hat{\boldsymbol{X}}\boldsymbol{W}_{M,v})) \tag{22}$$

$$p=Predictor(\boldsymbol{X};\theta_{pre}) \tag{23}$$

$$p_{m_1,M}=Predictor(\boldsymbol{X}_{m_1,M};\theta_{pre}) \tag{24}$$

$$\mathcal{L}_p=Loss(p,y)+\frac{\alpha_p}{|\mathcal{M}|}\sum_{m_1}Loss(p_{m_1,M},y) \tag{25}$$

where p is the final prediction, α_p is the weight of loss, and $\mathcal{L}_p$ is the predictive loss. Notably, here we calculate the predictive losses of cross-modal representations $\boldsymbol{X}_{m_1,M}$ to enhance their discriminative power. Similarly, we design multiple supervisory signals for the learning of inter-modal attention mechanisms to capture important inter-modal interactions and suppress useless and noisy interactions:

(1) Modality Mixing Operation: Drawing inspiration from the mix-up technique (Zhang et al., 2018; Verma et al., 2019) that mixes two samples to obtain an augmented sample, we design a modality mixing operation that generates mixed representations by weightedly mixing the representations of queries and keys. Since the query itself is used to generate the mixed representation, the mixed representation contains more information relevant to the query than the original key. Therefore, we adopt a ranking-based training strategy to ensure that the attention value between the query and the original key is less than that between the query and the mixed representation, enabling the model to identify more informative inter-modal interactions. Taking modality m_1 as the query and m_2 as the key and value, modality mixing operation is defined as:

$$\boldsymbol{K}^{mix}_{m_1,m_2}=\lambda\cdot\boldsymbol{X}_{m_2}+(1-\lambda)\cdot\boldsymbol{X}_{m_1} \tag{26}$$

$$\boldsymbol{A}^{mix}_{m_1,m_2}=\frac{\boldsymbol{X}_{m_1}\boldsymbol{W}_{m_1,q_M}(\boldsymbol{K}^{mix}_{m_1,m_2}\boldsymbol{W}_{m_1,k_M})^R}{\sqrt{d}} \tag{27}$$

where λ is a random value between 0 and 1, and $\boldsymbol{K}^{mix}_{m_1,m_2}$ denotes the mixed representation. The modality mixing loss is then defined as:

$$\mathcal{L}^{mix}_{m_1,m_2}=max((\boldsymbol{A}_{m_1,m_2}-\boldsymbol{A}^{mix}_{m_1,m_2})\cdot\mathbb{I}+\gamma_1,0) \tag{28}$$

where $\mathbb{I} \in \mathbb{R}^{T\times T}$ represents a diagonal matrix with values of 1 on the diagonal and 0 elsewhere. We add $\mathbb{I}$ to enable a more accurate and safer supervised learning, ensuring that the query features used for mixing and the query features used for attention come from the same time slice. $\mathcal{L}^{mix}_{m_1,m_2}$ encourages attention mechanisms to identify features that are more relevant to the target query.

(2) **Noise Addition** and **Variant Constraint**: These two techniques for inter-modal attention are consistent with those for intra-modal attention, and therefore the description is omitted here.

3.3 Model Optimization

We jointly optimize the predictive loss and the supervised losses for attention mechanisms:

$$\begin{aligned}\mathcal{L}=&\mathcal{L}_p+\alpha\cdot(\mathcal{L}^{de}+\mathcal{L}^{n})\\&+\alpha_{mix}\cdot\mathcal{L}^{mix}+\alpha_c\cdot(\mathcal{L}^{std}+\mathcal{L}^{con})\end{aligned} \tag{29}$$

where $\mathcal{L}$ is the final loss of the model, α, α_{mix} and α_c are the weights of the supervised losses. $\mathcal{L}^n$ is the sum of all intra- and inter-modal noise addition losses, and other losses vice verse.

4 Experiments

SAM-LML is evaluated on CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), UR-FUNNY (Hasan et al., 2019), and MUStARD (Castro et al., 2019) datasets. Due to space limitations, **the introductions of experimental settings, baselines, and datasets are shown in Appendix.**

Table 1: The results on the CMU-MOSI and CMU-MOSEI datasets. The results labeled with † are obtained from their papers, and other results are derived from our experiments. The best results are highlighted.

	CMU-MOSI					CMU-MOSEI				
	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
MFM (Tsai et al., 2019)	33.3	80.0	80.1	0.948	0.664	50.8	83.4	83.4	0.580	0.722
Self-MM (Yu et al., 2021)	45.8	84.9	84.8	0.731	0.785	53.0	85.2	85.2	0.540	0.763
ConFEDE† (Yang et al., 2023)	42.3	85.5	85.5	0.742	0.782	54.9	85.8	85.8	0.522	0.780
DMD + MCIS† (Yang et al., 2024a)	46.5	86.3	86.3	-	-	55.2	87.3	87.2	-	-
AtCAF† (Huang et al., 2025)	46.5	88.6	88.5	0.650	0.831	**55.9**	87.0	86.8	**0.508**	0.785
C-MIB (Mai et al., 2023c)	47.7	87.8	87.8	0.662	0.835	52.7	86.9	86.8	0.542	0.784
ITHP (Xiao et al., 2024)	47.7	88.5	88.5	0.663	0.856	52.2	87.1	87.1	0.550	0.792
Multimodal Boosting (Mai et al., 2024)	49.1	88.5	88.4	0.634	0.855	54.0	86.5	86.5	0.523	0.779
SAM-LML	**49.4**	**89.2**	**89.1**	**0.628**	**0.861**	55.0	**87.9**	**87.9**	0.516	**0.795**

Table 2: The comparison on UR-FUNNY.

Model	Accuracy	Number of Parameters
HKT‡ (Hasan et al., 2021)	77.4	-
DMD+SuCI‡ (Yang et al., 2024b)	70.8	-
AtCAF‡ (Huang et al., 2025)	72.1	-
HKT (Hasan et al., 2021)	76.5	17,066,564
MCL (Mai et al., 2023a)	77.7	**13,762,973**
MGCL (Mai et al., 2023b)	78.1	14,062,342
SAM-LML	**78.7**	14,809,023

Table 3: The results on MUStARD.

Model	Accuracy	Number of Parameters
HKT‡ (Hasan et al., 2021)	79.4	-
MO-Sarcation‡ (Tomar et al., 2023)	79.7	-
HKT (Hasan et al., 2021)	76.5	17,101,372
MCL (Mai et al., 2023a)	77.9	13,828,449
MGCL (Mai et al., 2023b)	77.9	14,282,000
SAM-LML	**80.9**	**13,589,833**

4.1 Performance Under Complete Input

The results on MSA are shown in Table 1. On CMU-MOSI, SAM-LML outperforms ITHP (Xiao et al., 2024) that also applies DeBERTa (He et al., 2021) as the language network by 1.7 points in Acc7 and 1.4 points in Acc2. On CMU-MOSEI, SAM-LML obtains the best results in Acc2, F1 score, and Corr.Generally, SAM-LML achieves state-of-the-art results on MSA. This is mainly because the proposed supervised attention mechanisms can highlight informative interactions and suppress noisy information, improving the robustness of fusion. Compared to Multimodal Boosting (Mai et al., 2024), SAM-LML demonstrates a considerable performance improvement (its unimodal networks are the same as ours to ensure a fair comparison). Multimodal Boosting transforms unimodal predictive losses to generate deterministic labels for unimodal contributions, which might be inaccurate as absolute labels of weights are hard to determine. In contrast, we use a ranking-based training strategy to encourage attention mechanisms to compare the relative importance of interactions, aligning well with the nature of attention: focusing on 'more' important information.

4.2 Results on MHD and MSD

To justify the generalizability of SAM-LML, we carry out experiments on the tasks of MHD and MSD on the UR-FUNNY (Hasan et al., 2019) and MUStARD (Castro et al., 2019) datasets. As shown in Table 2 and Table 3, SAM-LML outperforms state-of-the-art methods MGCL (Mai et al., 2023b) and MO-Sarcation (Tomar et al., 2023), respectively. Moreover, SAM-LML has moderate parameters compared to baselines. Therefore, SAM-LML obtains state-of-the-art results with moderate complexity, indicating the generalizability of SAM-LML with respect to other multimodal tasks.

4.3 Ablation Experiments

(1) The importance of supervised attention: As presented in Table 4, in the case of 'W/O Supervised Learning', the performance of SAM-LML decreases by over 3.5 points in Acc7 and Acc2. Moreover, when intra- or inter-modal attention has no supervisory signals, the performance also decreases, indicating both self-attention and cross-modal attention require explicit supervisory signals to enhance their performance. These results demonstrate the importance of supervised attention learning that can capture more informative interactions and highlight more discriminative representations.

(2) The importance of different components: When any component of SAM-LML is removed, the performance of SAM-LML deteriorates. Among them, the noise addition operation is the most crucial. This is because the noise addition operation is key to identifying noisy data and learning effective information from noisy interactions.The variance constraint is also relatively important as it enables the attention matrix to better distinguish the significance of different interactions, preventing all

Table 4: Ablation experiments on CMU-MOSI.

	Acc7	Acc2	F1	MAE	Corr
W/O Supervised Learning	45.8	85.6	85.4	0.680	0.821
Unsupervised Intra-Modal Attention	47.4	87.9	87.9	0.648	0.840
Unsupervised Inter-Modal Attention	46.6	89.0	89.0	0.666	0.837
W/O Noise Addition Operation	47.3	87.8	87.8	0.655	0.840
W/O Modality Mixing Operation	46.3	88.7	88.7	0.648	0.846
W/O Modality Decomposition Replacement	48.4	88.2	88.2	0.649	0.846
W/O Variance Constraint	47.7	88.1	88.1	0.664	0.841
W/O Context Constraint	48.1	**89.2**	**89.1**	0.659	0.849
SAM-LML	**49.4**	**89.2**	**89.1**	**0.628**	**0.861**

interactions from being assigned the same weight. In contrast, the context constraint is less important because it is only applied to the self-attention matrix and the constraint is relatively simple.

4.4 Performance Under Incomplete Inputs

As shown in Table 5, we evaluate the performance of SAM-LML under different modality missing rates (MR), with the missing rate ranging from mild (0.1) to severe (0.7). The baselines include MCTN (Pham et al., 2019), MMIN (Zhao et al., 2021), GCNet (Lian et al., 2023), IMDer (Wang et al., 2023), and MoMKE (Xu et al., 2024b). The results suggest that SAM-LML outperforms competitive baselines in the majority of missing settings on two widely-used datasets, highlighting its robustness under missing modalities. Moreover, the average performance (Avg) of SAM-LML across various missing rates consistently outperforms all baselines. Specifically, On CMU-MOSI, SAM-LML achieves average Acc2 and Acc7 of 74.6% and 35.9%, respectively, demonstrating improvements of 1.2 and 6.7 points over GCNet (Lian et al., 2023). On CMU-MOSEI, the average Acc2 and Acc7 of SAM-LML reach 80.0% and 48.2%, respectively, outperforming advanced baselines MoMKE (Xu et al., 2024b) and IMDer (Wang et al., 2023). We argue that this is mainly because SAM-LML explicitly simulates the absence of modalities via the noise addition operation, which enables the model to better adapt to scenarios with missing modalities and allows the model to focus more on discriminative information in the presence of missing or noisy modalities.

4.5 Discussion on Noisy Modalities

We mix all modalities with Gaussian noise during training and testing to evaluate the effectiveness of SAM-LML in handling noisy modalities. The noisy rate (NR) is set to 10% - 70%. For a comprehensive comparison, we present the results of C-MIB (Mai et al., 2023c) and Multimodal Boosting (Mai et al., 2024) that also address noisy modality issue. Following our baselines, we use Acc2 and MAE as evaluation metrics. As shown in Table 13, SAM-LML significantly outperforms competitive baselines in all settings (especially for MAE), and the performance of SAM-LML does not show a significant decrease even when NR is as high as 0.7. This is mainly because our supervised attention explicitly recognizes noisy information and interactions, and the noise addition operation in Eq. 5- 6 encourages augmented noisy features to be more informative than pure noise data, enabling the model to discover discriminative information within augmented noisy features. These results indicate the superiority and robustness of SAM-LML.

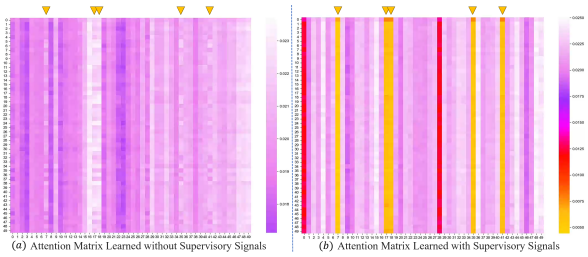


Figure 3: Visualization of attention matrices.

4.6 Visualization of Attention Matrices

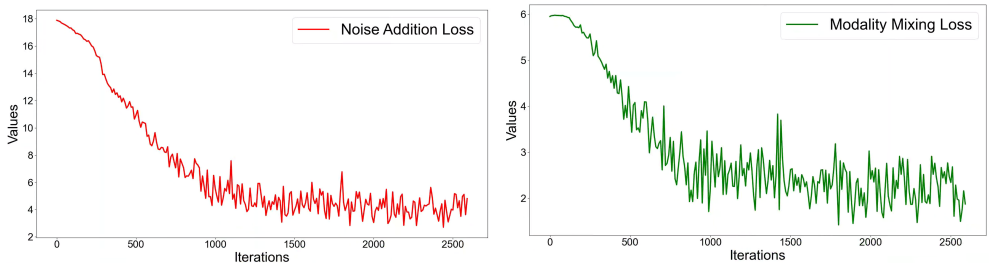
Figure 3 (b) displays an inter-modal attention matrix between language and visual modalities for a sample in CMU-MOSEI. For comparison, Figure 3 (a) visualizes the attention matrix learned without explicit supervisory signals. To assess whether SAM-LML can identify noisy interactions, we randomly replace the features of specific time slices in the language sequence with Gaussian noise (columns 7, 17, 18, 35, and 41). When supervised attention is applied, SAM-LML assigns lower weights to noisy information, suggesting that it can accurately identify noise. In contrast, when the attention mechanism is trained without explicit supervisory signals, the model sometimes fails to recognize noisy interactions, and even assigns higher weights to noisy interactions (columns 17 and 18), indicating attention mechanism learned without supervisory signals is prone to focusing on noisy interactions and often has poorer interpretability. Additionally, the values of the elements within the attention matrix learned with supervisory signals show greater differences, indicating stronger discriminability of the attention mechanism.

4.7 Learning Curves of Supervised Losses

Here we present the learning curves of the noise addition loss and modality mixing loss. As shown

Table 5: The results under incomplete modalities on the CMU-MOSI and CMU-MOSEI datasets.

Datasets	MR	MCTN	MMIN	GCNet	IMDer	MoMKE	SAM-LML
		Acc2 / F1 / Acc7	Acc2 / F1 / Acc7	Acc2 / F1 / Acc7	Acc2 / F1 / Acc7	Acc2 / F1 / Acc7	Acc2 / F1 / Acc7
MOSI	0.1	78.4 / 78.5 / 39.8	81.8 / 81.8 / 41.2	82.2 / 82.3 / 35.4	83.5 / 83.4 / 42.1	82.5 / 81.6 / 35.1	**84.8 / 84.7 / 45.6**
	0.2	75.6 / 75.7 / 38.5	79.0 / 79.1 / 38.9	79.4 / 79.5 / 34.6	80.5 / 80.5 /41.6	78.5 / 76.6 / 32.9	**81.2 / 81.2 / 42.9**
	0.3	71.3 / 71.2 / 35.5	76.1 / 76.2 / 36.9	77.1 / 77.2 / 32.5	77.4 / 77.6 / 37.4	74.4 / 71.7 / 30.6	**78.0 / 78.1 / 37.5**
	0.4	68.0 / 67.6 / 32.9	71.7 / 71.6 / 34.9	**75.3 / 75.4** / 30.3	66.5 / 66.3 / 35.2	70.7 / 67.5 / 28.4	74.8 / 74.7 / **37.8**
	0.5	65.4 / 64.8 / 31.2	67.2 / 66.5 / 34.2	**72.4 / 72.4** / 29.3	65.2 / 65.4 / 29.5	66.9 / 63.2 / 26.2	71.0 / 70.9 / **32.8**
	0.6	63.8 / 62.5 / **29.7**	64.9 / 64.0 / 29.1	64.3 / 64.5 / 23.6	66.0 / 65.5 / 27.0	63.2 / 58.9 / 23.9	**66.8 / 66.6** / 28.1
	0.7	61.2 / 59.0 / 27.5	62.8 / 61.0 / **28.4**	64.8 / 64.9 / 18.9	62.2 / 60.4 / 26.5	60.6 / 55.9 / 22.4	**65.9 / 65.6** / 26.4
	Avg.	69.1 / 68.5 / 33.6	71.9 / 71.5 / 34.5	73.6 / 73.7 / 29.2	71.6 / 71.3 / 34.2	71.0 / 67.9 / 28.5	**74.6 / 74.5 / 35.9**
MOSEI	0.1	81.8 / 81.6 / 49.8	81.9 / 81.3 / 50.6	84.3 / 84.5 / 46.9	82.9 / 82.9 / **52.1**	**85.1 / 84.7** / 47.2	84.7 / 84.5 / 51.9
	0.2	79.0 / 78.7 / 48.6	79.8 / 78.8 / 49.6	83.3 / 82.3 / 45.1	80.6 / 79.7 / 51.3	83.3 / 82.7 / 45.4	**83.8 / 83.7 / 51.7**
	0.3	76.9 / 76.2 / 47.4	77.2 / 75.5 / 48.1	81.2 / **81.5** / 44.5	78.7 / 77.8 / **49.6**	81.6 / 80.7 / 43.6	**81.9 / 81.6** / 48.7
	0.4	74.3 / 74.1 / 45.6	75.2 / 72.6 / 47.5	79.3 / 77.7 / 43.4	73.7 / 73.3 / 48.0	79.8 / 78.7 / 41.7	**80.1 / 79.5 / 48.3**
	0.5	73.6 / 72.6 / 45.1	73.9 / 70.7 / 46.7	77.3 / 74.7 / 41.8	72.1 / 68.4 / 46.6	**78.1** / 76.7 / 39.8	77.8 / **77.4 / 46.9**
	0.6	73.2 / 71.1 / 43.8	73.2 / 70.3 / **45.6**	75.9 / 73.2 / 38.6	70.8 / 65.9 / 45.0	76.3 / 74.7 / 37.9	**76.7 / 76.4 / 45.6**
	0.7	72.7 / 70.5 / 43.6	73.1 / 69.5 / **44.8**	75.0 / 73.7 / 38.1	69.1 / 66.6 / 44.1	**75.2** / 73.3 / 36.7	74.9 / **74.6** / 44.5
	Avg.	75.9 / 75.0 / 46.3	76.3 / 74.1 / 47.6	79.5 / 78.2 / 42.6	75.4 / 73.5 / 48.1	79.9 / 78.8 / 41.8	**80.0 / 79.7 / 48.2**



(a) Noise Addition Loss (b) Modality Mixing Loss

Figure 4: Learning curves of the proposed losses on the CMU-MOSI dataset. We average the data of every 10 iterations to make losses smoother.

Table 6: Discussion on noisy modalities on the CMU-MOSI and CMU-MOSEI datasets.

Datasets	NR	C-MIB	Multimodal Boosting	SAM-LML
		Acc2 / MAE	Acc2 / MAE	Acc2 / MAE
MOSI	0.1	87.8 / 0.670	86.7 / 0.678	**88.4 / 0.636**
	0.2	87.5 / 0.726	86.1 / 0.738	**88.1 / 0.665**
	0.3	86.4 / 0.912	86.4 / 0.785	**87.8 / 0.663**
	0.4	83.2 / 1.366	85.5 / 0.841	**87.6 / 0.666**
	0.5	84.9 / 1.660	86.1 / 1.172	**88.1 / 0.666**
	0.6	80.8 / 2.595	82.0 / 1.355	**87.5 / 0.660**
	0.7	82.1 / 3.146	84.4 / 1.750	**87.3 / 0.669**
	Avg.	84.7 / 1.582	85.3 / 1.046	**87.8 / 0.661**
MOSEI	0.1	86.1 / 0.545	86.4 / 0.544	**87.0 / 0.521**
	0.2	84.5 / 0.582	86.6 / 0.557	**87.0 / 0.525**
	0.3	85.6 / 0.622	85.5 / 0.623	**87.3 / 0.522**
	0.4	84.4 / 0.703	85.3 / 0.682	**87.2 / 0.525**
	0.5	83.7 / 0.875	84.1 / 0.724	**86.6 / 0.529**
	0.6	82.4 / 1.054	85.4 / 0.924	**87.0 / 0.532**
	0.7	80.5 / 1.404	80.3 / 1.125	**85.9 / 0.545**
	Avg.	83.9 / 0.826	84.8 / 0.740	**86.9 / 0.528**

in Fig. 4, both the losses decrease significantly as training deepens, indicating that they are optimized in the desired directions. Among them, the learning of modality mixing loss is relatively difficult, and the loss curve oscillates severely. This may be because in the modality mixing operation, the two inter-modal interactions used for comparison both contain rich modality correlation information, making them hard to distinguish. The noise addition loss decreases significantly, but the final loss does not approach 0. This is mainly because we impose an additional constraint on the noise addition operation, requiring the weights between augmented noisy features and the query to be greater than those between pure noise data and the query. This demands the model to extract useful information from augmented noisy features, which is challenging but beneficial for handling noisy modalities.

5 Conclusion

We propose a framework that jointly addresses missing and noisy modalities. Specifically, we design supervisory signals for the learning of attention weights to ensure that attention mechanisms can focus on discriminative information and suppress noisy information. We further propose a ranking-based optimization strategy that can avoid training noise caused by inaccurate absolute labels. Our SAM-LML obtains state-of-the-art results on multiple datasets under various settings.

6 Acknowledgment

This work was supported by the National Natural Science Foundation of China (62076262, 61673402, 61273270, 60802069), and a grant from South China Normal University (Grant No. 24KJ01).

Limitations

The limitation of this paper is that the operations of generating supervisory signals for attention weights increase the training time, which might be a potential drawback when scaling to multimodal large language models. Nevertheless, the inference time of SAM-LML will not increase as the axillary losses will not be calculated during inference (see Section A.8).

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629.

Thodsaporn Chay-intr, Yujun Chen, Kobkrit Viriyayudhakorn, and Thanaruk Theeramunkong. 2025. Llavac: Fine-tuning llava as a multimodal sentiment classifier. *arXiv preprint arXiv:2502.02938*.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep: A collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964.

Yuanyue Deng, Jintang Bian, Shisong Wu, Jianhuang Lai, and Xiaohua Xie. 2025. Multiplex graph aggregation and feature refinement for unsupervised incomplete multimodal emotion recognition. *Information Fusion*, 114:102711.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*.

Dimitris Gkoumas, Qiuchi Li, C. Lioma, Yijun Yu, and Da wei Song. 2021. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197.

Peizhu Gong, Jin Liu, Xiliang Zhang, Xingye Li, Lai Wei, and Huihua He. 2025. Adaptive multimodal graph integration network for multimodal sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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)*, pages 1726–1736.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12972–12980.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056.

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. 2022. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–696.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Ddecoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. 2025. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9215–9223.

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

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

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

Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Learning representations from imperfect time series data via tensor rank regularization. In *ACL*, pages 1569–1576.

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

Ronghao Lin and Haifeng Hu. 2024. Adapt and explore: Multimodal mixup for representation learning. *Information Fusion*, 105:102216.

Rui Liu, Haolin Zuo, Zheng Lian, Björn W Schuller, and Haizhou Li. 2024. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sijie Mai, Ya Sun, Aolin Xiong, Ying Zeng, and Haifeng Hu. 2024. Multimodal boosting: Addressing noisy modalities and identifying modality contribution. *IEEE Transactions on Multimedia*, 26:3018–3033.

Sijie Mai, Ya Sun, Ying Zeng, and Haifeng Hu. 2023a. Excavating multimodal correlation for representation learning. *Information Fusion*, 91:542–555.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2023b. Learning from the global view: Supervised contrastive learning of multimodal representation. *Information Fusion*, 100:101920.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2023c. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2023d. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289.

Huisheng Mao, Baozheng Zhang, Hua Xu, Ziqi Yuan, and Yihe Liu. 2023. Robust-msa: Understanding the impact of modality noise on multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16458–16460.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6892–6899.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. 2022. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, pages 17782–17800. PMLR.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Shuwei Qian and Chongjun Wang. 2023. Com: Contrastive masked-attention model for incomplete multimodal learning. *Neural Networks*, 162:443–455.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and E. Hoque. 2020. Integrating multimodal information in large pretrained transformers. *ACL*, 2020:2359–2369.

Yuge Shi, Brooks Paige, Philip Torr, and 1 others. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8992–8999.

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

Mohit Tomar, Abhisek Tiwari, Tulika Saha, and Sriparna Saha. 2023. Your tone speaks louder than your face! modality order infused multi-modal sarcasm detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3926–3933.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414.

Yao Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. In *ICLR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR.

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

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

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. 2024a. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446.

Wenxin Xu, Hexin Jiang, and 1 others. 2024b. Leveraging knowledge of modality experts for incomplete multimodal learning. In *ACM Multimedia 2024*.

Xiaojun Xue, Chunxia Zhang, Zhendong Niu, and Xindong Wu. 2023. Multi-level attention map network for multimodal sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):5105–5118.

Zihui Xue and Radu Marculescu. 2023. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2575–2584.

Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. 2024a. Towards multimodal sentiment analysis debiasing via bias purification. In *European Conference on Computer Vision*, pages 464–481. Springer.

Dingkang Yang, Dongling Xiao, Ke Li, Yuzheng Wang, Zhaoyu Chen, Jinjie Wei, and Lihua Zhang. 2024b. Towards multimodal human intention understanding debiasing via subject-deconfounding. *arXiv preprint arXiv:2403.05025*.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, pages 2236–2246.

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

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. 2022. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 2924–2934.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27456–27466.

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 (Volume 1: Long Papers)*, pages 2608–2618.

A Appendix

A.1 Unimodal Networks

In this section, we outline the structures of unimodal networks and illustrate how to generate unimodal representations for later fusion. Following the state-of-the-art methods (Xiao et al., 2024; Hasan et al., 2021), we use pre-trained language models (He et al., 2021; Lan et al., 2020) to generate high-level language features. For all tasks, the procedures of the language network are as follows:

$$\begin{aligned} \hat{\boldsymbol{U}}_l &= \text{PLM}(\boldsymbol{Z}_l;\ \theta_l) \in \mathbb{R}^{T\times d_l} \\ \boldsymbol{U}_l &= (\hat{\boldsymbol{U}}_l \boldsymbol{W}_{pro} + \boldsymbol{b}_{pro}) \in \mathbb{R}^{T\times d} \end{aligned} \tag{30}$$

where PLM represents the pre-trained language model, $\boldsymbol{Z}_l$ denotes the input token sequence and T denotes the sequence length. $\boldsymbol{W}_{pro} \in \mathbb{R}^{d_l\times d}$ and $\boldsymbol{b}_{pro} \in \mathbb{R}^{1\times d}$ are trainable parameters that transform the output dimensionality of the language network to match the pre-defined shared feature dimensionality d. For the task of MSA, the procedures for the acoustic and visual networks, which use transformer encoders (Vaswani et al., 2017), are shown as below ($m \in \{a, v\}$):

$$\begin{aligned} \hat{\boldsymbol{U}}_m &= \text{Conv 1D}(\boldsymbol{Z}_m;\ K_m) \in \mathbb{R}^{T\times d} \\ \boldsymbol{U}_m &= \text{Transformer}(\hat{\boldsymbol{U}}_m;\ \theta_m) \in \mathbb{R}^{T\times d} \end{aligned} \tag{31}$$

where Conv 1D is the temporal convolution with kernel size K_m being 3. The obtained unimodal representation $\boldsymbol{U}_m$ is used for attention learning.

Please note that for MHD and MSD, to effectively identify humor-related information, following prior methods (Hasan et al., 2021; Mai et al., 2023b), Human Centric Feature (HCF) is additionally extracted from the language modality to serve as the fourth modality (see (Hasan et al., 2021) for more details) and is denoted as $\boldsymbol{Z}_h \in \mathbb{R}^{T\times d_h}$. Furthermore, each sample includes a target punchline segment along with its preceding context segments. By concatenating the feature sequences of the punchline and context segments in the temporal dimension, we obtain the unimodal inputs $\boldsymbol{Z}_m \in \mathbb{R}^{T\times d_m}$ ($m \in \mathcal{M} = \{a, v, l, h\}$). The unimodal network for the HCF modality, which also consists of transformer encoders, has a structure similar to those for the visual and acoustic modalities. Specifically, for MHD and MSD, the procedures of the transformer-based unimodal networks are outlined as follows ($m \in \{a, v, h\}$):

$$\begin{aligned} \hat{\boldsymbol{U}}_m &= \text{Transformer}(\boldsymbol{Z}_m;\ \theta_m) \in \mathbb{R}^{T\times d_m} \\ \boldsymbol{U}_m &= \text{Conv 1D}(\hat{\boldsymbol{U}}_m;\ K_m) \in \mathbb{R}^{T\times d} \end{aligned} \tag{32}$$

Moreover, to simplify the subsequent modeling process, we merge the language and HCF modalities using a straightforward linear layer:

$$\boldsymbol{U}_l \longleftarrow \text{Linear}(\boldsymbol{U}_l \oplus \boldsymbol{U}_h;\ \theta_{lin}) \in \mathbb{R}^{T\times d} \tag{33}$$

A.2 Dataset Composition

We use the following datasets to evaluate the performance of SAM-LML:

(1) **CMU-MOSI** (Zadeh et al., 2016): CMU-MOSI is a widely-used dataset for MSA that consists of more than 2,000 video segments sourced from the Internet. Each video segment is manually labeled with sentiment intensity on a scale from -3 to 3, where 3 denotes the strongest positive sentiment and -3 denotes the strongest negative sentiment.

(2) **CMU-MOSEI** (Zadeh et al., 2018): CMU-MOSEI is a large-scale MSA dataset with over 22,000 video segments from over 1,000 YouTube speakers across 250 varied topics. These segments are selected at random from a wide variety of topics and solo video presentations. Each video segment is annotated with two kinds of labels: emotions that are divided into six different classes and sentiment scores that range from -3 to 3. To assess the performance of SAM-LML on the MSA task, we employ the sentiment labels from the CMU-MOSEI dataset, which align with the sentiment scale of the CMU-MOSI dataset.

(3) **UR-FUNNY** (Hasan et al., 2019): Designed for the task of MHD, this dataset comprises TED talk videos with 1,741 speakers. Each target video segment in the dataset, termed a 'punchline', encompasses language, acoustic, and visual modalities. The segments preceding these punchlines, known as context segments, are input into the model alongside the punchlines for contextual analysis. The punchlines are identified using the 'laughter' tag in the transcripts, which indicates when the

audience laughed during the talk. Negative samples are identified in a similar manner, with target punchline segments not followed by the 'laughter' tag. UR-FUNNY comprises 7,614 training, 980 validation, and 994 testing samples. To be consistent with state-of-the-art methods (Hasan et al., 2021; Mai et al., 2023a,b), we utilize version 2 of the UR-FUNNY dataset to evaluate the proposed model.

(4) **MUStARD** (Castro et al., 2019): MUStARD is a sarcasm detection dataset derived from well-known TV shows including Friends, The Big Bang Theory, The Golden Girls, and Sarcasmaholics. The dataset includes 690 video segments, each manually labeled as sarcastic or non-sarcastic. In addition to the punchline segments, MUStARD also contains the preceding conversations (context segments) for each punchline to provide contextual information.

A.3 Assessment Criteria

For MSA, we evaluate the performance of SAM-LML and baselines using the following assessment criteria: (1) **Acc7**: Acc7 measures the model's ability to accurately categorize sentiment scores into seven specific classes. For calculating Acc7, predictions and labels are rounded to the nearest integer between -3 and 3; (2) **Acc2**: Acc2 measures the model's ability to accurately distinguish positive and negative sentiments in a binary classification setting; (3) **F1 score**: It is a metric that averages precision and recall for binary sentiment classification task. Neutral segments are discarded when calculating both the Acc2 and the F1 score; (4) **MAE**: the mean absolute error between model predictions and annotated labels; (5) **Corr**: the correlation coefficient measuring the strength and direction of alignment between model predictions and annotated labels.

For MHD and MSD, in alignment with prior methodologies (Hasan et al., 2021; Mai et al., 2023a,b), we report the binary accuracy (i.e., humorous or non-humorous, sarcastic or non-sarcastic) of the model.

A.4 Feature Extraction Strategy

We use the following tools to extract the features of the modalities:

(1) Visual Modality: For the MSA task, following prior works (Mai et al., 2023d; Xiao et al., 2024), we use Facet[1] to extract visual features such as facial action units, facial landmarks, and head positioning, forming a temporal sequence that captures facial expressions and body gestures over time. For the MHD and MSD tasks, to be consistent with state-of-the-art algorithms (Hasan et al., 2021; Mai et al., 2023a), we adopt OpenFace 2 (Baltrusaitis et al., 2018) to extract facial action units, rigid and non-rigid facial shape parameters, etc.

(2) Acoustic Modality: We use COVAREP (Degottex et al., 2014) to extract time-series acoustic features, including 12 Mel-frequency cepstral coefficients, pitch tracking, speech polarity, glottal closure instants, and spectral envelope, etc. Extracted over the full audio of each segment, these features capture dynamic variations in vocal tone throughout the speech.

(3) Language Modality: For the MSA task, following the state-of-the-art methods (Xiao et al., 2024), DeBERTa (He et al., 2021) is employed to extract high-level textual representations. For the MHD and MSD tasks, following the state-of-the-art methods (Hasan et al., 2021; Mai et al., 2023a), ALBERT (Lan et al., 2020) is applied as the language network. Notably, for MHD and MSD, we concatenate the punchline and context token sequences to generate the final input of the language network: $\boldsymbol{U}_l = \boldsymbol{C}_l \oplus [SEP] \oplus \boldsymbol{P}_l$, where the $[SEP]$ token is used to separate the context tokens $\boldsymbol{C}_l$ from the punchline tokens $\boldsymbol{P}_l$ (Hasan et al., 2021).

For the CMU-MOSI dataset, the dimensionality of language, acoustic, and visual features are 768, 74, and 47, respectively. For the CMU-MOSEI dataset, the dimensionality of the corresponding unimodal features are 768, 74, and 35, respectively. For the UR-FUNNY and MUStARD datasets, the feature dimensionality of language, acoustic, visual, and HCF modalities are 768, 60, 36, and 4 respectively. For the feature extraction details of the HCF modality, please refer to (Hasan et al., 2021).

A.5 Experimental Details

(1) **Hyper-parameter setting**: We implement the proposed SAM-LML using the PyTorch framework on an NVIDIA RTX2080Ti GPU with CUDA version 11.6 and PyTorch version 1.13.1, using the AdamW optimizer (Loshchilov and Hutter, 2019). Hyperparameter settings are detailed in Table 7.

[1] iMotions 2017. https://imotions.com/

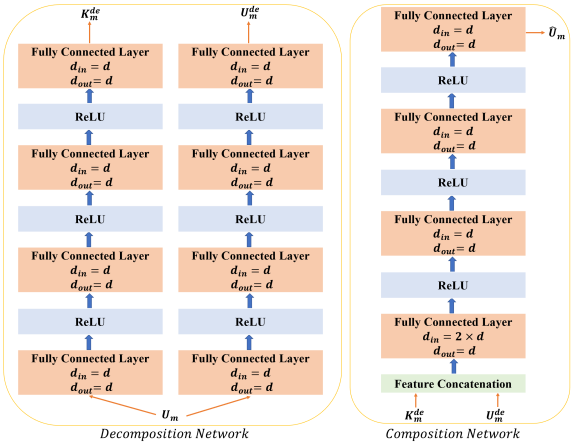


Figure 5: The structures of composition and decomposition networks in the modality decomposition replacement operation.

Following previous work (Gkoumas et al., 2021), we perform a random grid search with 50 random iterations on the validation set to find optimal hyperparameters, and save the best-performing hyperparameter configuration. After hyperparameter search, we retrain the model for five times with the optimal settings under different random seeds, and report the mean results of the five-time running. The structures of composition and decomposition networks in the modality decomposition replacement operation are shown in Figure 5.

(2) **Protocol for the evaluation of missing modalities**: For the evaluation of missing modalities, we comprehensively evaluate the effectiveness of various methods on multimodal datasets under different modality missing rates. The missing rate is defined as:

$$MR = 1 - \frac{\sum_{i=1}^{N} M_i}{N \times |\mathcal{M}|} \tag{34}$$

where M_i denotes the number of available modalities in the i-th sample, N is the total number of samples, and $|\mathcal{M}|$ is the number of modalities for the task. For each sample with $|\mathcal{M}|$ modalities, we randomly mask a subset of modalities with a probability corresponding to the missing rate MR, while ensuring that at least one modality is retained for each sample, i.e., $M_i \geq 1$. This constraint guarantees that the missing rate does not exceed $\frac{|\mathcal{M}|-1}{|\mathcal{M}|}$. In our experiments, when the number of modalities is $|\mathcal{M}|= 3$, we select te MR from the set $\{0.1, 0.2, \ldots, 0.7\}$. Particularly, when $MR = 0.7$, each multimodal sample randomly retains one modality, which is the most severe modality missing scenario. To make a fair comparison with baselines, the same missing rate configuration is applied across the training, validation, and test phases, following the protocol adopted in previous works (Wang et al., 2023; Lian et al., 2023).

(3) **Protocol for the evaluation of noisy modalities**: For the evaluation of noisy modalities, we generate the noisy features using the following equation:

$$\boldsymbol{U}_m^n = (1 - NR) \cdot \boldsymbol{U}_m + NR \cdot \mathcal{N} \tag{35}$$

where $\boldsymbol{U}_m$ is the unimodal representation, NR is the noisy rate ranging from 0.1 to 0.7, $\mathcal{N}$ is the Gaussian noise data of mean 0 and variance 1, and $\boldsymbol{U}_m^n$ is the noisy representation that is used in the attention mechanisms. Notably, for all modalities of all samples, we apply the aforementioned noise mixing operation to simulate noise in the real world. Since noise at the input level ultimately affects the features, it is acceptable to apply noise at the feature level. Moreover, for the sake of fairness and privacy protection, different algorithms for multimodal affective computing typically model based on the same set of features, making it more feasible to apply noise at the feature level.

A.6 Baselines

The compared baselines for MSA include:

(1) **Multimodal Factorization Model** (**MFM**) (Tsai et al., 2019): MFM decomposes multimodal features into two independent factor sets: multimodal discriminative factors and modality-specific generative factors. This decomposition facilitates learning effective multimodal representations, and the factorized features can be used to understand critical inter-modal interactions for multimodal learning.

(2) **Contrastive FEature DEcomposition** (**ConFEDE**) (Yang et al., 2023): ConFEDE concurrently engages in contrastive representation learning and feature decomposition, thereby enhancing the multimodal representation.

(3) **Information-Theoretic Hierarchical Perception** (**ITHP**) (Xiao et al., 2024): Based on the information bottleneck principle, ITHP designates one core modality and regards the remaining modalities as detectors within the information pathway that serve to distill information flow.

(4) **Decoupled Multimodal Distillation** (**DMD**) (Li et al., 2023): To enhance the discriminative features of each modality, DMD facilitates flexible and adaptive cross-modal knowledge distillation,

Table 7: Hyperparameter Settings of SAM-LML. MSE and BCE denotes mean square error and binary cross-entropy, respectively.

	CMU-MOSI	CMU-MOSEI	UR-FUNNY	MUStARD
Loss Function	MSE	MSE	BCE	BCE
Batch Size	48	64	64	48
Learning Rate	1e-5	1e-5	1e-6	5e-6
Shared Dimensionality d	120	180	150	48
Diagonal Elements ρ	5	5	6	7
Variance Margin β	0.1	0.1	0.1	0.1
Context Margin γ	0.5	0.5	0.5	1
Margin γ_1	1	1	0.3	0.8
Predictive Loss Weight α_p	1	0.1	1e-5	0.05
Decouple Loss Weight α_{de}	0.01	0.5	1e-4	0.01
Loss Weight α	0.005	0.005	1e-4	1e-4
Loss Weight α_{mix}	0.01	0.01	0.1	1e-5
Loss Weight α_c	0.005	0.005	1e-5	0.05

which decouples each unimodal representation into modality-irrelevant/-exclusive spaces and employs a graph distillation unit to process each decoupled part in a more specialized and effective manner.

(5) **Multimodal Boosting** (Mai et al., 2024): It employs multiple base learners, where different base learners focus on different aspects of multimodal learning. To assess individual contributions, Multimodal Boosting introduces a contribution learning module that dynamically determines each base learner's contribution and the noisy level of unimodal representations.

(6) **Complete Multimodal Information Bottleneck** (**C-MIB**) (Mai et al., 2023c): It uses the information bottleneck principle to reduce redundancy and noise in unimodal and multimodal features, which is use for the baseline of noisy modality handling.

(7) **Self-Supervised Multi-task Multimodal sentiment analysis network** (**Self-MM**) (Yu et al., 2021): It calculates sentiment labels of individual modalities by leveraging the global labels of multimodal samples in a self-supervised manner, thereby extracting more discriminative unimodal features.

(8) **Multimodal Counterfactual Inference Sentiment** (**MCIS**) (Yang et al., 2024a): MCIS is a framework based on causality to identify harmful biases in pre-trained models and generate unbiased decisions from biased observations by comparing factual and counterfactual outcomes.

(9) **Attention-based Causality-Aware Fusion** (**AtCAF**) (Huang et al., 2025): AtCAF uses a counterfactual cross-modal attention module to capture causal relationships in training data, constructing a comprehensive causality chain to trace causal trajectories from user inputs to model outputs.

(10) **Missing Modality Imagination Network (MMIN)** (Zhao et al., 2021): MMIN generates robust joint multimodal representations via cross-modal imagination, enabling the prediction of any missing modality from the available ones under diverse missing modality conditions.

(11) **Multimodal Cyclic Translation Network (MCTN)** (Pham et al., 2019): MCTN proposes a translation-based approach to learn robust joint representations by translating between each two modalities, enabling sentiment prediction using only one source modality during testing.

(12) **Graph Complete Network (GCNet)** (Lian et al., 2023): GCNet tackles the missing modality issue in conversational scenarios by introducing Speaker GNN and Temporal GNN to model speaker and temporal dependencies, and jointly optimizes classification and reconstruction tasks to utilize both complete and incomplete modality data.

(13) **Incomplete Multimodality-Diffused emotion recognition (IMDer)** (Wang et al., 2023): IMDer introduces a score-based diffusion model to recover missing modalities by transforming Gaussian noise into modality-specific distributions, and leverages available modalities as conditional guidance to ensure consistency and semantic alignment during recovery.

(14) **Mixture of Modality Knowledge Experts (MoMKE)** (Xu et al., 2024b): MoMKE uses a two-

stage framework: unimodal experts are first trained independently, then jointly trained with combined unimodal and joint representations. A Soft Router is introduced to dynamically mix these representations to enable enriched and adaptive multimodal learning under incomplete modality scenarios.

The additional baselines for the MHD and MSD tasks include:

(1) **Multimodal Adaptation Gate ALBERT (MAG-ALBERT)** (Rahman et al., 2020): MAG-ALBERT incorporates a multimodal adaptation gate, allowing large pre-trained transformers to handle multimodal data during fine-tuning.

(2) **Multimodal Global Contrastive Learning (MGCL)** (Mai et al., 2023b): MGCL conducts supervised contrastive learning on multimodal representations and devises various operations to generate positive and negative samples for each representation.

(3) **Multimodal Correlation Learning (MCL)** (Mai et al., 2023a): MCL designs supervised multimodal correlation learning task to retain modality-specific information and acquire a more discriminative embedding space.

(4) **Subject Causal Intervention (SuCI)** (Yang et al., 2024b): It proposes a simple and effective causal intervention module to disentangle the impact of subjects as unobserved confounders and achieve unbiased predictions via true causal effects.

(5) **Humor Knowledge Enriched Transformer (HKT)** (Hasan et al., 2021): HKT is a promising method for MHD and MSD, utilizing humor centric feature as external knowledge to address ambiguity and sentiment information hidden in the language modality.

(6) **Modality Order-driven module for Sarcasm detection (MO-Sarcation)** (Tomar et al., 2023): It introduces a modality order-driven fusion module that is integrated into a transformer network, which can fuse modalities in an ordered way.

A.7 Additional Visualizations for Attention Matrices

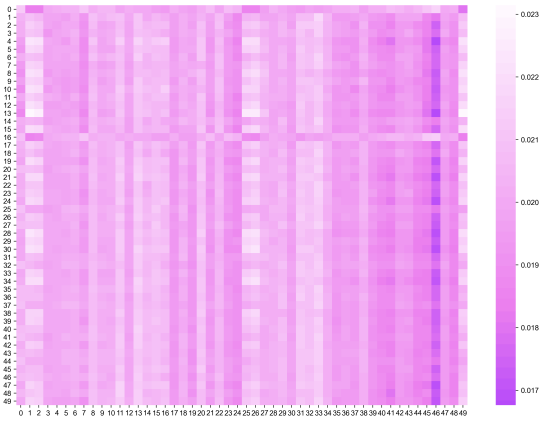
Figure 3 displays the visual-language attention matrix where the features of specific time slices in the language sequence are replaced with Gaussian noise. In this section, we visualize the language-acoustic attention matrix and language-visual attention matrix as well as their counterparts learned without supervisory signals. As illustrated in Figure 6, similar to the case in Figure 3, when the features of some specific time slices in the visual/acoustic sequence are replaced with Gaussian noise, the proposed attention mechanism can accurately identify the noise and assign very low attention values to noisy interactions. In contrast, when the attention mechanism is trained without explicit supervisory signals, the model fails to recognize noisy interactions, and even assigns higher weights to noisy interactions, indicating attention mechanism learned without supervisory signals is prone to focusing on noisy interactions and often has poorer interpretability. In fact, we found that the aforementioned problem is more severe when visual/acoustic features are replaced by noise, where the attention mechanism trained without explicit supervisory signals tends to highlight all noisy interactions. This is because, compared to the language modality, visual/acoustic features contain less discriminative information, making it difficult for conventional attention mechanisms to distinguish between visual/acoustic features and noisy features. These findings further underscore the necessity of designing supervisory signals for attention mechanisms to enable them to more effectively learn the discriminability of different interactions and information.

Table 8: The running time of SAM-LML for one epoch on the CMU-MOSEI dataset.

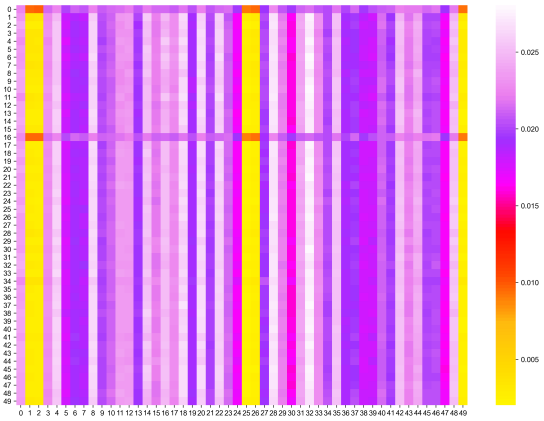
	Training Time (s)	Testing Time (s)
ITHP (Xiao et al., 2024)	111.1	6.9
SAM-LML (W/O Supervision)	**109.8**	**6.7**
SAM-LML	144.5	**6.7**

A.8 Running Time Analysis

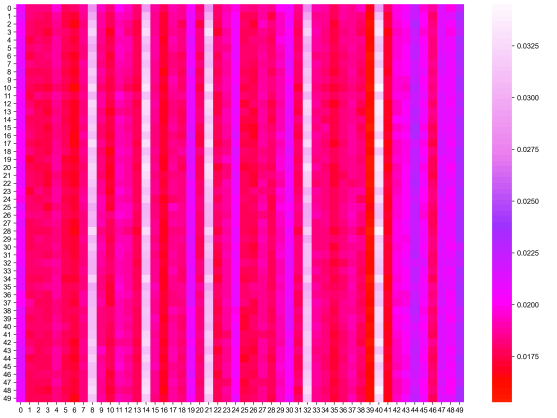
In this section, we examine the computational efficiency of SAM-LML. The introduction of multiple supervisory signals to optimize the attention mechanisms leads to an increase in the running time of SAM-LML. However, by employing a simple model architecture, we manage to mitigate the complexity of SAM-LML, thereby offsetting the extended running time to some extent. As demonstrated in Table 8, SAM-LML requires 144.5 seconds for a training epoch and 6.7 seconds for a testing epoch on the CMU-MOSEI dataset, while the state-of-the-art model ITHP takes 111.1 seconds for training and 6.9 seconds for testing (under identical conditions). These results suggest that while the training time of SAM-LML increases, its testing time experiences a slight reduction. Furthermore, eliminating the supervision on atten-



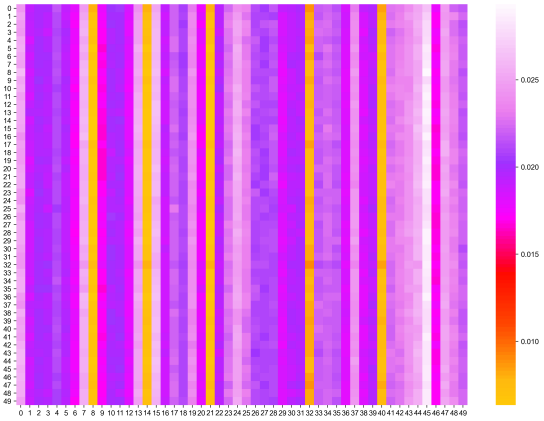
(a) Language-Acoustic Attention Matrix (W/O Supervisory Signals)



(b) Language-Acoustic Attention Matrix



(c) Language-Visual Attention Matrix (W/O Supervisory Signals)



(d) Language-Visual Attention Matrix

Figure 6: Additional visualizations for attention matrices.

tion mechanisms shaves off 34.7 seconds per training epoch for SAM-LML, which is approximately 24.0% of its initial training time. Note that the number of parameters of SAM-LML remains acceptable as we only introduce modality decomposition and composition networks in our supervised attention mechanisms (see Table 2 and Table 3).

Table 9: GFLOPs comparison on UR-FUNNY and MUSTARD datasets.

	UR-FUNNY	MUSTARD
Metric	GFLOPs	GFLOPs
HKT	5.27	6.36
MGCL	5.07	6.14
SAM-LML	**5.01**	**6.07**

A.9 Model Complexity Analysis

In this section, we further test the FLOPs of the models on the MHD and MSD tasks to evaluate the time complexity. As shown in Table 9, the FLOPs of SAM-LML are fewer than the baselines in both datasets. Therefore, we believe that the computational overhead of SAM-LML is acceptable.

A.10 Significant Test

In this section, we present the results of a significant test conducted between SAM-LML and current state-of-the-art (SOTA) approaches, utilizing the t-test method. The SOTA1 and SOTA2 for the MSA task are ITHP (Xiao et al., 2024) and Multimodal Boosting (Mai et al., 2024) respectively, while for the MSD and MHD tasks are HKT (Hasan et al., 2021) and MGCL (Mai et al., 2023b) respectively. To make a comprehensive comparison, each model is executed ten times with distinct random seeds. As illustrated in Table 10, the t-test results indicate that there is a statistically significant divergence between SAM-LML and the SOTA methods across the majority of the assessment metrics (where p<0.05 signifies a statistically significant difference). This suggests that the enhancements made by the proposed SAM-LML are acceptable.

A.11 Discussion on Unaligned Data

In this section, we evaluate the proposed SAM-LML and previous SOTA methods on the unaligned data. As shown in Table 11, generally there is a slight decrease in the performance when the data become unaligned, partly due to the excessively long acoustic and visual sequences (as acoustic and visual modalities contain relatively less discriminative information compared to the language modality). However, the performance of SAM-LML is still better than competitive baselines. This is mainly because the proposed supervised attention mechanism can effectively handle arbitrary-length sequences and accurately mine the correlation between these sequences even if the sequences are extremely long. Thus, unaligned data do not have a significant impact on the performance of SAM-LML.

A.12 Importance of Ranking-based Training Strategy

In this section, we analyze the benefits of ranking-based training strategy by reformatting it as deterministic training strategy that directly assigns absolute labels for weights. In deterministic setting, for the modified noise addition loss (NAL-v2), we set the attention labels for noisy interactions to zero after softmax. For the modified modality mixing loss (MML-v2), we set a large label (0.1) for attention values between the query and mixed features after softmax. For the modified modality decomposition replacement loss (MDRL-v2), we set a very small label (1e-4) for attention values between the query and $\boldsymbol{K}_m^{de}$ after softmax.

It can be inferred from Table 12 that converting ranking-based losses to deterministic forms leads to performance degradation. It is mainly because ranking-based losses avoid inaccurate learning and guide the model to compare importance, aligning with the nature of attention: focusing on more informative interactions.

A.13 Hyperparameter Analysis

In this section, we conduct the experiments for some important hyperparameters to analyze how they influence the robustness and performance of SAM-LML on the CMU-MOSI dataset, where α, α_c, and α_{mix} are the weights for modality mixing loss, context and variance constraints, and the combination of noise addition and modality decomposition replacement losses, respectively. As presented in Fig. 7 (a), (b), and (c), for all weights, the optimal performance is achieved when they are moderate values. This is because when the weights are too small, the effect of losses is not significant enough, and when they are too large, they can interfere with the training of the main task.

Moreover, we have also analyzed γ_1, which is a hyperparameter that determines the minimal mar-

Table 10: Paired t-test analysis between SAM-LML and state-of-the-art baselines.

	CMU-MOSI					CMU-MOSEI					UR-Funny	MUStARD
	Acc7	Acc2	F1	MAE	Corr	Acc7	Acc2	F1	MAE	Corr	Acc	Acc
SOTA1	3.68e-4	0.008	0.007	1.73e-4	0.452	1.17e-4	9.82e-4	8.57e-4	8.33e-5	0.362	0.003	0.007
SOTA2	0.035	0.006	0.006	0.052	0.398	0.009	6.24e-4	1.93e-4	0.022	0.003	0.017	0.008

Table 11: Performance comparison of different models on CMU-MOSI and CMU-MOSEI datasets.

Model	Data	CMU-MOSI			CMU-MOSEI		
		Acc7	Acc2	MAE	Acc7	Acc2	MAE
C-MIB	aligned	47.7	87.8	0.662	52.7	86.9	0.542
	unaligned	47.4	87.5	0.655	52.2	86.7	0.539
Multimodal Boosting	aligned	49.1	88.5	0.634	54.0	86.5	0.523
	unaligned	47.9	88.1	0.642	53.7	86.8	0.529
SAM-LML	aligned	**49.4**	89.2	**0.628**	55.0	**87.9**	**0.516**
	unaligned	49.1	**89.5**	0.623	**55.2**	87.8	0.518

Table 12: Performance comparison on CMU-MOSI and CMU-MOSEI datasets.

	MOSI		MOSEI	
Model	Acc7	Acc2	Acc7	Acc2
NAL-v2	48.6	88.6	54.5	87.4
MML-v2	47.4	88.5	54.0	87.2
MDRL-v2	47.9	88.3	53.9	87.0
SAM-LML	**49.4**	**89.2**	**55.0**	**87.9**

gin between two weights. As shown in Fig. 7 (d), γ_1 should be a large value so that we can increase the distinguishability between different weights and enhance the attention mechanism's ability to differentiate interactions of varying degrees of importance. Furthermore, we find that when the hyperparameters are set to specific values (e.g., when γ_1 is 2), the performance is better than that obtained by grid search, indicating the performance of SAM-LML can be further improved after a more careful hyperparameter tuning.

A.14 Additional Results on Other Noise

In addition to Gaussian noise, we evaluate the robustness of SAM-LML against Laplace noise and random erasing noise (randomly select a portion of the features and set them to zero to simulate data loss). For each sample, we randomly select one type of noise to add to the features, and the results are shown in Table 13. The noisy rate (NR) is set to 10% - 70%. For a comprehensive comparison, we present the results of C-MIB (Mai et al., 2023c) and Multimodal Boosting (Mai et al., 2024), which use the same training and testing settings as ours. It can be seen from Table 13 that SAM-LML still

Table 13: Additional results on other types of noise on the CMU-MOSI and CMU-MOSEI datasets.

	NR	C-MIB	Multimodal Boosting	SAM-LML
		Acc2 / MAE	Acc2 / MAE	Acc2 / MAE
MOSI	0.1	87.6 / 0.681	87.3/0.660	**88.1 / 0.647**
	0.2	87.3 / 0.695	85.8/0.726	**87.8 / 0.661**
	0.3	85.0 / 0.890	85.6/0.757	**87.8 / 0.649**
	0.4	83.7 / 1.019	87.6/0.798	**87.9 / 0.659**
	0.5	81.5 / 1.629	86.9/0.839	**87.6 / 0.648**
	0.6	79.4 / 1.274	84.6/1.048	**87.6 / 0.656**
	0.7	81.4 / 3.443	85.4/1.432	**87.4 / 0.673**
	Avg	83.7 / 1.376	86.2/0.894	**87.7 / 0.656**
MOSEI	0.1	86.8 / 0.531	86.8 / 0.555	**87.3 / 0.525**
	0.2	85.9 / 0.587	86.0 / 0.581	**87.1 / 0.530**
	0.3	85.7 / 0.606	85.2 / 0.586	**86.8 / 0.532**
	0.4	84.0 / 0.646	85.4 / 0.680	**86.9 / 0.525**
	0.5	83.9 / 0.798	85.0 / 0.746	**86.7 / 0.527**
	0.6	82.2 / 1.046	86.0 / 0.957	**86.5 / 0.526**
	0.7	77.6 / 1.271	81.2 / 1.018	**86.5 / 0.529**
	Avg	83.7 / 0.784	85.1 / 0.732	**86.8 / 0.528**

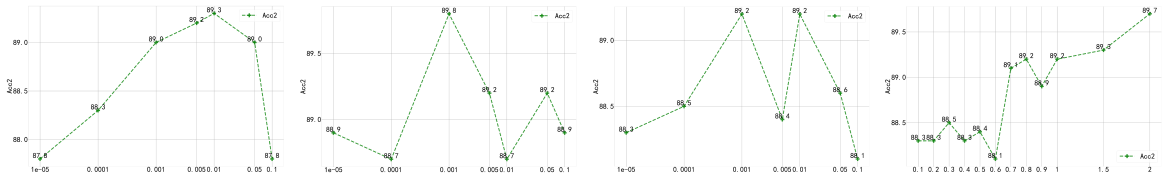
significantly outperforms competitive baselines under other noises (especially when the noise rate is large), demonstrating the effectiveness and generalization ability of SAM-LML.

Table 14: Performance comparison of different models across modalities.

Modality	Model	Acc7	Acc2	MAE
Acoustic	HuBERT	**49.7**	**89.6**	0.630
	Covarep (Default)	49.4	89.2	**0.628**
Visual	CLIP	**49.4**	**89.5**	**0.623**
	Facet (Default)	**49.4**	89.2	0.628
Language	DeBERTa-v3-large	**50.6**	**90.5**	**0.597**
	DeBERTa-v3-base (Default)	49.4	89.2	0.628

A.15 Discussion on Unimodal Models

In this section, we conduct experiments to evaluate the performance of SAM-LML under various unimodal model. For acoustic modality, we change the default COVAREP with more powerful HuBERT (Hsu et al., 2021). For visual modal-



(a) Loss Weight α (b) Loss Weight α_c (c) Loss Weight α_{mix} (d) Margin γ_1

Figure 7: Model performance w.r.t the change of hyperparameters.

ity, we compare the popular CLIP (Radford et al., 2021) with the default Facet. For language modality, we change the default DeBERTa-v3-base with more powerful DeBERTa-v3-large. As shown in Table 14, using more advanced language network leads to more significant performance improvement. Moreover, using more advanced acoustic and visual models generally brings improvement to the multimodal framework. Specifically, HuBERT and CLIP achieve the best performance in acoustic and visual feature extraction, respectively, as they are both highly expressive pre-trained models trained on large-scale data. Nevertheless, the performance improvement is not as significant as the change of language model, further indicating the dominant role of the language modality in multimodal affective computing.