KHaR: Knowledge Harvesting and Refinement for multimodal sentiment analysis

Anonymous ACL submission

Abstract

Multimodal sentiment analysis deduces a user's sentiment by integrating information from different modalities. Previous methods mainly focus on using complex fusion networks to learn effective joint embeddings and achieved significant improvement, while ignoring refined processing within modalities, resulting in models that achieve suboptimal results on more delicate sentiment analysis. In this paper, we propose a novel framework KHaR, which harvests rich intra-modality knowledge through domain-specific adapters and utilizes mixture of experts to refine the knowledge to capture more detailed intra-modality information. In addition, we design a contrastive learning to further explore the information correlation between samples of similar sentiment intensity. To fuse the extracted features effectively, we employ the multimodal information bottleneck to filter out irrelevant information and retain the most salient features for sentiment analysis. Extensive experiments show that KHaR achieves superior performance on four benchmark datasets, and especially achieves significant improvement on the more refined sentiment analysis(e.g.Acc-7, Acc-5).

1 Introduction

With the rapid progress of deep learning, significant breakthroughs have been made in the field of MSA. Contemporary studies in MSA predominantly concentrate on developing advanced fusion methodologies to effectively integrate cross-modal information. The simplest approach, concatenation or early fusion, directly stacked the unimodal representations from different modalities(Poria et al., 2016; Kampman et al., 2018; Wu and Liang, 2010). Outer product-based fusion captures pairwise feature correlations at the cost of quadratic computational complexity(Zadeh et al., 2017). Attention-based fusion mechanisms dynamically weight the contribution of each modality, allowing for more flexible and adaptive integration(Tsai et al., 2019; Zhang et al., 2023; Feng et al., 2024; Sun and Tian, 2025). Graph-based fusion methods model the relationships between different modalities as a graph, enabling more complex interactions and dependencies(Huang et al., 2024; Lin et al., 2022).

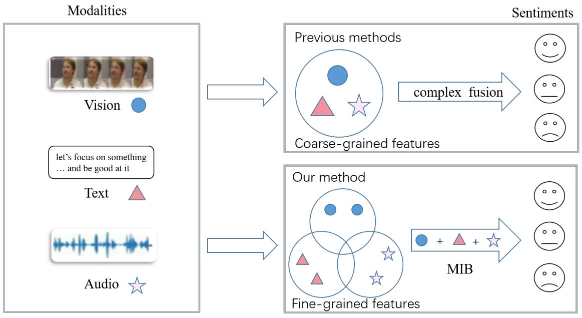


Figure 1: The previous methods and our method.

Although these methods are quite successful, they neglect fine-grained intra-modal representation learning. Most methods rely on coarse-grained features, such as using the first token extracted by BERT to represent the text modality(Devlin et al., 2019; Hazarika et al., 2020; Niu et al., 2021). Averaging of all tokens or max pooling are used to represent the audio and video modalities(Tsai et al., 2019; Zong et al., 2023). As a result, when evaluated on fine-grained sentiment tasks (e.g.Acc-7, Acc-5), such approaches yield suboptimal performance as they fail to capture subtle but decisive sentiment signals in individual modalities before fusion. For example, subtle changes in mouth movement or body language in a video can significantly enhance sentiment analysis(Luo et al., 2021; Hjelm et al., 2018). Figure 1 illustrates the difference between previous methods and our method.

To overcome these limitations, we propose a novel framework KHaR (Knowledge Harvesting and Refinement), which learns noise-robust representations by knowledge harvesting and refinement.

First, domain-specific adapters are injected into BERT/Transformer encoders to extract modality-aware features(textual semantic, acoustic patterns, and facial dynamics). These features are then processed by a Mixture of Experts(MoE) layer that dynamically activates domain-specialized MLPs to refine modality-specific information. Crucially, we design a threshold-based contrastive learning mechanism where samples with sentiment intensity difference $\leq \kappa$(empirically set to 0.4) form positive pairs, while others are treated as negative pairs, enabling fine-grained discrimination between similar emotions(e.g., κ=0.4 groups y=1.0 with y=1.3 but separates from y=2.0). Finally, we employ the multimodal information bottleneck to compress cross-modal features by theoretically eliminating redundant inter-modal correlations while preserving the sentiment-discriminative patterns for compact and sufficient fusion.

The main contributions of our work can be summarized as follows:

- We propose the KHaR framework to systematically address fine-grained intra-modality feature extraction, which ensures the preservation and enhancement of subtle emotional signals before fusion through knowledge harvesting from domain adapters and refinement of MoE.
- A dynamic contrastive mechanism is designed to align cross-modal features based on continuous sentiment intensity levels, significantly improving fine-grained emotion differentiation compared to traditional binary polarity-based approaches.
- Empirical evaluations across four benchmark datasets consistently show KHaR outperforming current state-of-the-art approaches. Ablation studies further confirmed the need for each component.

2 Related Work

In this part, we review the previous works in the field of MSA, focusing on two main aspects: multimodal sentiment analysis and contrastive learning.

2.1 Multimodal Sentiment Analysis

Previous research on MSA mainly focuses on intra-modality representation learning and inter-modality fusion. For intra-modality representation learning methods, Mai et al. (2020) and Hazarika et al. (2020) utilized adversarial learning to learn the common and specific features of each modality. Zhang et al. (2023) designed an Adaptive Hyper-modality Learning(AHL) module that, with the aid of language features at various scales, learns to create a representation capable of suppressing irrelevancies and conflicts present in visual and audio features. For inter-modality fusion methods, Zadeh et al. (2017) proposed a tensor fusion network method to effectively capture and analyze the interactions between different modalities by computing the cartesian product. Feng et al. (2024) proposed a dynamic attention fusion method based on the contribution of each modality. Unlike the previous works, we extract more detailed representation information within modalities by harvesting with refinement.

2.2 Contrastive Learning

Contrastive learning learns discriminative representations by pulling positive pairs closer while pushing away negative pairs. Recent researches on contrastive learning in multimodal sentiment analysis are mainly focused on two aspects: supervised contrastive learning(Zha et al., 2023; Mai et al., 2022; Yang et al., 2024; Khosla et al., 2020) and self-supervised contrastive learning(Akbari et al., 2021; Chen et al., 2020; He et al., 2020). The core difference between them is whether to use label information to construct positive and negative sample pairs. For example,Chen et al. (2020) proposed a self-supervised contrastive learning framework, SimCLR, which learns representations by maximizing agreement between augmented views of the same image. Khosla et al. (2020) introduced a supervised contrastive loss, which exploits the similarity of label information to maximize the consistency of samples in the same class. Inspired by Yang et al. (2024), we propose a dynamic contrastive learning mechanism that aligns cross-modal features based on the sentiment intensity levels, significantly improving fine-grained emotion differentiation compared to traditional binary polarity-based approaches.

3 Methodology

3.1 Overall Architecture

As shown in Figure 2, the KHaR framework first extracts unimodal features from raw video, text, and audio inputs. An adapter and encoder jointly learn modality-specific representations enriched with affective knowledge. A MoE module then dynami-

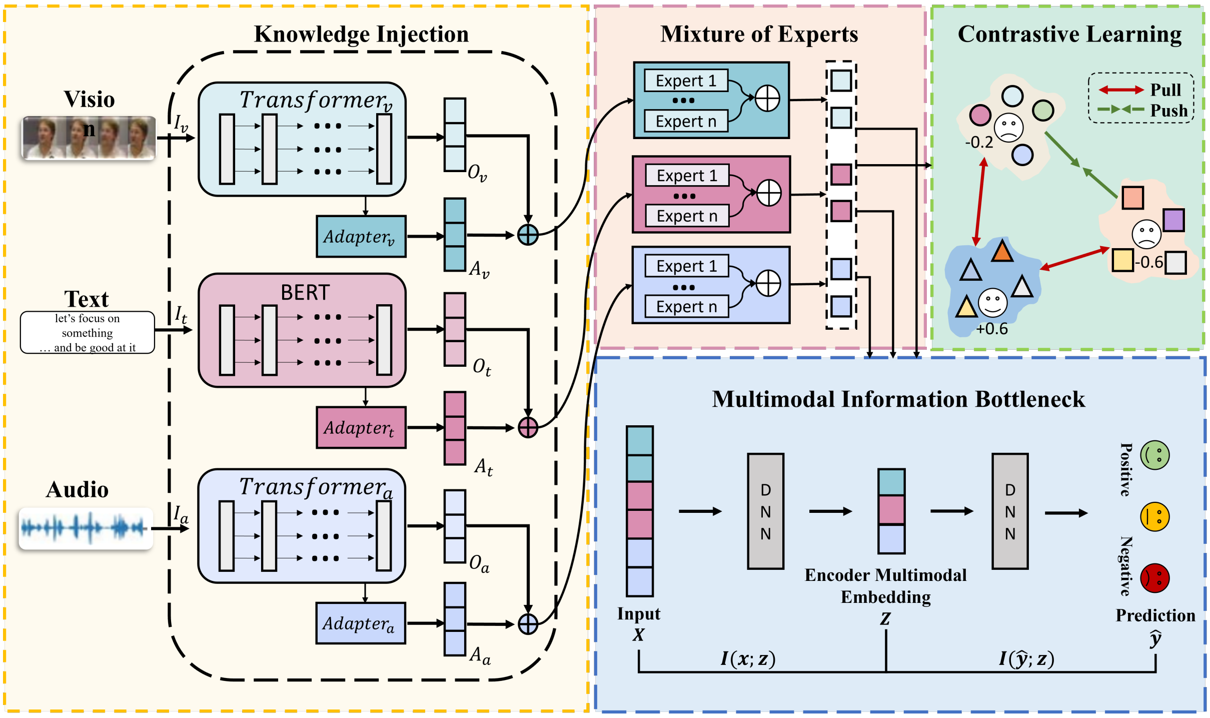


Figure 2: The overall architecture of KHaR. The blue, pink and purple modules represent the relevant operations of vision, text and audio.

cally selects domain-specific experts(each of which is an MLP)to compress and refine critical features. To capture emotional differences across modalities, a dynamic contrastive learning mechanism adjusts the similarity distribution accordingly. Finally, an information bottleneck module constrains the fusion process by filtering out irrelevant information while preserving sentiment-relevant cues.

3.2 Problem Definition

MSA task typically processes visual, textual, and acoustic data streams as input features. The sequences of three modalities are represented as triplet (I_v, I_t, I_a), which include $I_v \in \mathbb{R}^{T_v \times d_v}$, $I_t \in \mathbb{R}^{T_t \times d_t}$, and $I_a \in \mathbb{R}^{T_a \times d_a}$, where $T_m, m \in \{v, t, a\}$ is the corresponding modal sequence length, d_m represents the vector dimension. Our main task is to accurately predict the sentiment intensity through feature extraction and fusion.

3.3 Knowledge of Harvest

In order to harvest a large amount of knowledge, we use encoders and adapters to encode each modality input $I_{m \in \{v,t,a\}}$, and obtain two types of representations: a comprehensive-level representation $O_m \in \mathbb{R}^{T_m \times d_m}$ and the knowledge-enhanced affective representations $A_m \in \mathbb{R}^{T_m \times d_m}$.

Comprehensive-level Representation. To obtain hierarchical semantic features and a solid foundation for subsequent sentiment-aware knowledge injection via the adapter, we extract comprehensive-level representations O_m from each modality m. For text input I_t, we employ a pretrained BERT encoder, while for visual and audio inputs I_v and I_a, we use modality-specific stacked Transformer encoders (Vaswani et al., 2017):

$$O_m, H_m = \text{Encoder}_m(I_m; \theta_m^{\text{encoder}}) \qquad (1)$$

where O_m denotes the final output of the encoder, serving as the comprehensive-level representation, and H_m represents the intermediate hidden states that capture modality-specific general or commonsense information.

Knowledge-Enhanced Affective Representation. To obtain richer affective feature information, we introduce external Adapter modules in the encoder layers to inject affective knowledge into the modality-specific representations.

As illustrated in Figure 3, we insert stacked Adapter blocks into the encoder architecture of visual, text, and audio modalities to construct the knowledge-enhanced affective representation A_m:

$$A_m = \text{Adapter}_m(I_m, H_m; \theta_m^{\text{adapter}}) \in \mathbb{R}^{T_m \times d_m} \qquad (2)$$

where $\theta_m^{\text{adapter}}$ denotes the pretrained parameters of the Adapter module corresponding to modality.

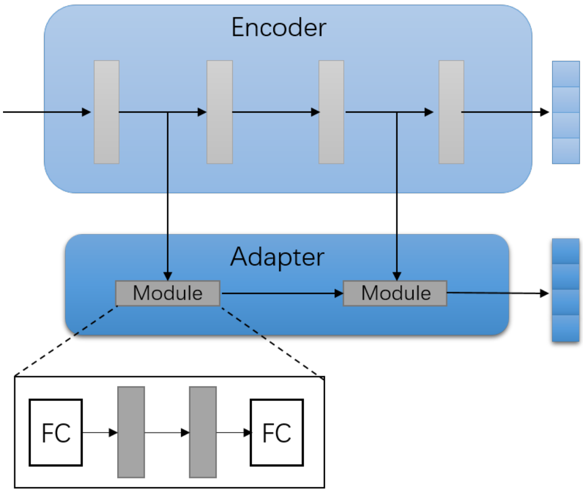


Figure 3: The architecture of BERT and Transformer encoders, Adapters, and their interconnections.

Through this design,we combine the above two representations to obtain the knowledge-enhanced representation X_m:

$$X_m = [A_m; H_m] \in \mathbb{R}^{T_m \times 2d_m} \quad (3)$$

3.4 Knowledge Refinement

To achieve fine-grained processing of widely harvested multimodal knowledge, KHaR adopts a modality-specific MoE layer, where each expert is an MLP network optimized for discriminative subfeatures. Gated network mechanisms dynamically activate specialized experts, for example, in text, different experts may focus on sentiment phrases, negation, or syntactic structures; In the video, experts locate local cues such as microexpressions or lip movements; While in audio, pitch change experts or spectrum pattern experts are selectively involved. This architecture ensures that only the most relevant experts deal with the nuances of a particular pattern, allowing the model to extract knowledge with high accuracy from a wide and noisy feature space.

Specifically, for each modality m, we construct an independent MoE layer comprising s expert networks $\{E_1, E_2, ..., E_s\}$ and a gating network G. All MoE layers across modalities share the same number of experts to ensure structural consistency. Given the unimodal feature X_m.The gating weights are computed as:

$$G(X_m) = \text{Softmax}\left(\text{KeepTopK}(W_g X_m, k)\right) \quad (4)$$

where $W_g \in \mathbb{R}^{d \times s}$ is a training weight parameter matrix, d is the modal feature dimension, and KeepTopK$(\cdot)$ denotes the sparse operation that retains the top k gating weights for each input. The final gated representation $\hat{X}_m \in \mathbb{R}^d$ is obtained by weighted aggregation of expert outputs:

$$\hat{X}_m = \sum_{i=1}^{s} G(X_m)_i \cdot E_i(X_m) \quad (5)$$

where $G(X_m)_i$ is the gating weight of the i-th expert and $E_i(X_m)$ is the output of the corresponding expert network.

To mitigate expert collapse—where a few experts dominate the activation distribution—we introduce a regularization term $\mathcal{L}^m_{\text{MoE}}(X_m)$ to encourage balanced expert utilization:

$$\mathcal{L}^m_{\text{MoE}}(X_m) = \omega \cdot \left(\text{CV}(\text{I}(X_m))^2 + \text{CV}(\text{L}(X_m))^2\right) \quad (6)$$

where CV$(\cdot)$ denotes the coefficient of variation, I$(\cdot)$ measures the overall activation importance of each expert across the batch, L$(\cdot)$ quantifies the expert usage frequency, and ω is a hyperparameter controlling the strength of the regularization.

3.5 Multimodal Fusion

In information theory, Mutual Information (MI) is used to measure the degree of interdependence between two random variables.Specifically, considering two variables x and y with marginal distributions $p(x)$ and $p(y)$ and joint distribution $p(x, y)$, MI is defined as the Kullback-Leibler (KL) divergence between the product of the marginal and joint distributions. The formal definition of MI is:

$$\begin{aligned} I(x; y) &= D_{\text{KL}}\left(p(x, y) \,\|\, p(x)p(y)\right) \\ &= \int dx\, dy\, p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \mathbb{E}_{(x,y)\sim p(x,y)}\left[\log \frac{p(x, y)}{p(x)p(y)}\right] \end{aligned} \quad (7)$$

To extend this principle to multimodal scenarios, we design a multimodal information bottleneck (MIB) module that refines the fused representation derived from independent single-modality encoders. The MIB module is designed to: (1) Redundancy filtering: suppress cross-modal redundancy; (2) Discriminative preservation: preserve discriminative cues necessary for prediction.

In the MIB module, we introduce the intermediate variable z and construct this objective function:

$$L_{MIB} = I(y; z) - \beta I(x; z) \quad (8)$$

where $I(\cdot;\cdot)$ denotes MI, and the parameter $\beta \geq 0$ determines the intensity of the compression. Our approach focuses on maximizing the objective function.

Since MI itself is difficult to calculate directly, we usually use variational methods to estimate it. Regarding the first term $I(y;z)$, we introduce an auxiliary distribution $q(y \mid z)$ to approximate the genuine posterior distribution $p(y \mid z)$, thereby obtaining its lower bound:

$$I(y;z) \geq \int dxdydzp(z \mid x)p(y \mid x)p(x)\log q(y \mid z) \tag{9}$$

Similarly, for the second term $I(x;z)$, we use the approximate distribution $q(z \mid x)$ to estimate the genuine posterior distribution $p(z \mid x)$, and obtain its upper bound form:

$$I(x;z) \leq \int dxdzp(x)p(z \mid x)\log \frac{p(z \mid x)}{q(z)} \tag{10}$$

where $q(z)$ is a variational approximation to the marginal distribution $p(z)$ which is often fixed to a standard normal Gaussian distribution.Combining these two estimates, we obtain the variational lower bound of the objective function:

$$\begin{aligned} L_{MIB} &= I(y;z) - \beta I(x;z) \geq J_{MIB} \\ &= \mathbb{E}\left[\log q(y \mid z) - \beta \cdot KL\left(p(z \mid x) \,\|\, q(z)\right)\right] \end{aligned} \tag{11}$$

where the expectation $\mathbb{E}$ is taken over $(x,y) \sim p(x,y)$ and $z \sim p(z \mid x)$. J_{MIB} is a variational lower bound of L_{MIB}. See the Appendix C for the specific formula derivation. Finally, by maximizing the objective function L_{MIB}, we equivalently minimize the loss $\mathcal{L}_{\text{MIB}}$, which is defined as follows:

$$\mathcal{L}_{\text{MIB}} = \mathbb{E}\left[\beta \cdot KL\left(p(z \mid x) \,\|\, q(z)\right) - \log q(y \mid z)\right] \tag{12}$$

3.6 Dynamic Contrastive Learning Based on Sentiment Intensity

Traditional contrastive learning methods (e.g., SimCLR or supervised contrastive learning) primarily focus on broad category-level separation, grouping all positive sentiment samples (sentiment intensity > 0) together and all negative ones (< 0) apart, while maximizing the distance between positive and negative representations. However, such methods often ignore the continuity and nuance of sentiment intensity, which may lead to semantic distortion. For example, traditional methods will classify samples with sentiment intensity +3 and sentiment intensity +0.1 as positive sample pairs, and samples with sentiment intensity +0.1 and sentiment intensity -0.1 as negative sample pairs, depending on whether the sentiment intensity is both positive and negative, which leads us to ignore the more detailed relationship between sentiment intensity. As a result, the previous studies(Lin et al., 2022; Yu et al., 2023) have achieved suboptimal results on more difficult multi-classification tasks.

Inspired by the previous work(Yang et al., 2024), we introduce a dynamic contrastive learning strategy that adapts sample pairing based on their sentiment intensity differences, thereby preserving the semantic continuity inherent to sentiment tasks. Specifically, for any two samples i and j in the batch, we define their emotion intensity gap as:

$$G(i,j) = |y_i - y_j|, \quad j \in \text{batch},\ j \neq i \tag{13}$$

where y_i and y_j denote the actual sentiment intensity of sample i and j, respectively. Considering a threshold κ(set to 0.4), sample pairs are categorized as follows:

$$\begin{aligned} G(i,j) \leq \kappa \Rightarrow (i,j) \in \text{pos pairs} \\ G(i,j) > \kappa \Rightarrow (i,j) \in \text{neg pairs} \end{aligned} \tag{14}$$

Based on the above steps, for sample i, the intermodal positive and negative sample pairs can be obtained as follows:

$$\begin{aligned} P^i_{inter} =& \{(V^i,T^i),(V^i,A^i),(T^i,A^i)\} \cup \\ & \{(V^i,T^j),(V^i,A^j),(T^i,A^j), \\ & (T^i,V^j),(A^i,V^j),(A^i,T^j) \\ & |(i,j) \in \text{pos pairs}\} \end{aligned} \tag{15}$$

$$\begin{aligned} N^i_{inter} =& \{(V^i,T^k),(V^i,A^k),(T^i,A^k), \\ & (T^i,V^k),(A^i,V^k),(A^i,T^k) \\ & |(i,k) \in \text{neg pairs}\} \end{aligned} \tag{16}$$

The intra-modal positive and negative sample pairs can be obtained as follows:

$$\begin{aligned} P^i_{intra} =& \{(V^i,V^j),(T^i,T^j),(A^i,A^j), \\ & |(i,j) \in \text{pos pairs}\} \end{aligned} \tag{17}$$

$$\begin{aligned} N^i_{intra} =& \{(V^i,V^k),(T^i,T^k),(A^i,A^k), \\ & |(i,j) \in \text{neg pairs}\} \end{aligned} \tag{18}$$

Then, we can get the P^i and N^i as follows:

$$P^i = P^i_{inter} \cup P^i_{intra} \tag{19}$$

$$N^i = N^i_{inter} \cup N^i_{intra} \tag{20}$$

For each anchor sample a, the model is trained to minimize the distance to positive samples and

Methods	CH-SIMS						CH-SIMSv2					
	MAE	Corr	Acc-5	Acc-3	Acc-2	F1	MAE	Corr	Acc-5	Acc-3	Acc-2	F1
TFN	0.432	0.591	39.30	65.12	78.38	78.62	0.303	0.707	52.55	72.21	80.14	80.14
LMF	0.441	0.576	40.53	64.68	77.77	77.88	0.367	0.557	47.79	64.90	74.18	73.88
MulT	0.453	0.564	37.94	64.77	78.56	79.66	0.291	**0.738**	54.81	73.19	80.68	80.73
BBFN*	0.430	0.564	40.92	61.05	78.12	77.88	0.300	0.708	53.29	71.47	78.53	78.41
Self-MM	0.425	0.595	41.53	65.47	80.04	**80.44**	0.311	0.695	52.77	72.61	79.69	79.76
CubeMLP*	0.419	0.593	41.79	**65.86**	77.68	77.59	0.334	0.648	52.90	71.95	78.53	78.53
CENet	0.471	0.534	33.92	62.58	77.90	77.53	0.310	0.699	53.04	73.10	79.56	79.63
TETFN	0.420	0.577	41.79	63.24	**81.18**	80.24	0.310	0.695	54.47	73.65	79.73	79.81
ALMT*	**0.408**	0.594	43.11	**65.86**	78.77	78.71	0.308	0.700	52.90	71.86	79.59	79.51
TMBL*	0.429	0.592	41.58	65.43	79.12	78.75	0.313	0.706	52.03	73.02	80.46	80.36
KHaR	**0.408**	**0.622**	**45.30**	65.43	79.43	79.52	**0.287**	0.735	**56.29**	**73.89**	**80.85**	**80.78**

Table 1: Results on CH-SIMS and CH-SIMSv2 datasets. The best result is in bold; * means the results are from (Feng et al., 2024), while other results are from (Mao et al., 2022).

Methods	MOSI						MOSEI					
	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1
TFN†	0.947	0.673	34.46	-	77.99/79.08	77.95/79.11	0.572	0.714	51.60	-	78.50/81.89	78.96/81.74
LMF†	0.950	0.651	33.82	-	77.90/79.18	77.80/79.15	0.576	0.717	51.59	-	80.54/83.48	80.94/83.36
MAG-BERT	0.727	0.781	43.62	-	82.37/84.43	82.50/84.61	0.543	0.755	52.67	-	82.51/84.82	82.77/84.71
BBFN*	0.796	0.744	43.88	-	80.32/82.47	80.21/82.44	0.545	0.760	52.88	-	82.87/85.73	83.13/85.56
MMIM*	0.744	0.780	44.75	-	82.51/84.30	82.38/84.23	0.550	0.761	51.88	-	**83.75**/85.42	**83.93**/85.26
CubeMLP*	0.755	0.772	43.44	-	80.76/82.32	81.77/84.23	0.537	0.761	53.35	-	82.36/85.23	82.61/85.04
PS-Mixer	0.794	0.748	44.31	-	80.3/82.1	80.3/82.1	0.537	0.765	53.0	-	83.1/86.1	83.10/86.10
ALMT*	0.712	0.793	46.79	-	83.97/85.82	84.05/85.86	0.530	0.774	53.62	-	81.54/85.99	81.05/86.05
SFTTR	0.709	0.795	46.5	-	82.94/84.6	82.92/84.63	0.536	0.772	53.7	-	82.89/85.99	83.15/85.92
MulT†	-	-	-	42.68	- / -	- / -	-	-	-	54.18	- / -	- / -
MISA†	-	-	-	47.08	- / -	- / -	-	-	-	53.63	- / -	- / -
Self-MM†	-	-	-	53.47	- / -	- / -	-	-	-	55.53	- / -	- / -
KHaR	**0.698**	**0.8**	**49.13**	**55.39**	**84.45/86.13**	**84.48/86.15**	**0.526**	**0.777**	**54.43**	**56.30**	83.37/**86.49**	83.02/**86.60**

Table 2: Results on MOSI and MOSEI datasets. The best result is in bold; * means the results are from (Feng et al., 2024), † means the results are from (Mao et al., 2022).

maximize the distance to negative samples. The contrastive learning loss is formulated as:

$$\mathcal{L}_{\text{CL}} = -\mathbb{E}_{i \in B} \log \left(\frac{\sum_{(a,p) \in P^i} \delta(a,p)}{\sum_{(a,q) \in P^i \cup N^i} \delta(a,q)} \right) \tag{21}$$

where $\delta(a,p) = exp(\frac{sim(a,p)}{\tau})$,which is used to meature the similarity between the anchor sample a and the positive sample p.

3.7 Optimization Objective

To achieve effective multimodal fusion and compact representation learning, the overall training objective of KHaR is formulated as the minimization of the following total loss function:

$$\mathcal{L} = \mathcal{L}_{\text{MoE}} + \mathcal{L}_{\text{MIB}} + \lambda \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{Task}} \tag{22}$$

where λ is a hyperparameter that controls the strength of contrastive learning, $\mathcal{L}_{\text{Task}}$ is to minimize the MAE between the predicted sentiment intensity $\hat{y}$ and the ground truth y.

4 Experiments

4.1 Datasets and Metrics

To fully evaluate the performance of KHaR, we employ four benchmark datasets: CMU-MOSI(Zadeh et al., 2016), CMU-MOSEI(Zadeh et al., 2018), CH-SIMS(Yu et al., 2020) and CH-SIMSV2(Liu et al., 2022). Appendices A and B describe the details of these datasets and the experimental setup.

For a more comprehensive comparsion, we divide our reporting results into regression and classification. For regression, we use the mean absolute error (MAE) and Pearson correlation (Corr) as evaluation metrics. For classification, we use the multi-class accuracy and F1-score as evaluation metrics. we report the accuracy of 2-class (Acc-2), 3-class (Acc-3) and 5-class (Acc-5) for CH-SIMS and CH-SIMSV2, and the accuracy of 2-class (Acc-2), 5-class (Acc-5) and 7-class (Acc-7) for MOSI and MOSEI. In addition, for MOSI and MOSEI, the Acc-2 and F1 score are computed for negative/positive(excluding 0) and negative/non-negative(including 0) classes.

4.2 Baselines

We compare the KHaR framework with state-of-the-art multimodal sentiment analysis methods, including: TFN(Zadeh et al., 2017), LMF(Liu et al., 2018), MulT(Tsai et al., 2019), MISA(Hazarika et al., 2020), MAG-BERT(Rahman et al., 2020), Self-MM(Yu et al., 2021), MMIM(Han et al., 2021b), BBFN(Han et al., 2021a), CENet(Wang et al., 2022), CubeMLP(Sun et al., 2022), PS-Mixer(Lin et al., 2023), TETFN(Wang et al., 2023), ALMT(Zhang et al., 2023), TMBL(Huang et al., 2024), and SFTTR(Sun and Tian, 2025).

4.3 Performance Comparison

We evaluate KHaR on four benchmark datasets and compare it with state-of-the-art methods. The results are shown in Table 1 and Table 2.

As shown in Table 1, for the Chinese datasets CH-SIMS and CH-SIMSv2, KHaR achieves the best results on most evaluation metrics. For example, on the CH-SIMSv2 dataset, compared with the MuLT model, KHaR achieves 1.37% improvement on the regression task MAE, and 2.7% improvement on the multi-classification task Acc-5, indicating that KHaR can extract effective feature information on complex datasets and make more accurate actual predictions.

For the English datasets MOSI and MOSEI, KHaR also achieves good experimental results. On the MOSI dataset, our model achieves the best results on all evaluation metrics. Compared with the second best result, KHaR improves by 1.55% in MAE and 0.62% in Corr. It is worth noting that on the more difficult and more detailed classification task Acc-7, our model achieves a significant improvement with a relative improvement of 5%. Similarly, on the MOSEI dataset, we achieve a relative improvement of 1.36% on the Acc-7 classification task, indicating that our model has a significant improvement in fine-grained sentim capture.

4.4 Ablation Study and Analysis

To validate the effectiveness of different components in our proposed KHaR model, we conduct comprehensive ablation studies on two benchmark datasets: CH-SIMS and MOSI. The evaluations cover the role of each modality, the impact of the MoE module, the contribution of contrastive learning, and the effects of MIB strategies. All results are reported in terms of MAE, correlation, Acc-5 for CH-SIMS and Acc-7 for MOSI.

Methods	CH-SIMS			MOSI		
	MAE	Corr	Acc-5	MAE	Corr	Acc-7
Role of Each Modality						
V+T	0.420	0.571	42.64	0.714	0.799	47.23
V+A	0.587	0.209	21.23	1.366	0.208	19.68
A+T	0.424	0.556	43.98	0.718	0.783	47.96
V+A+T	**0.408**	**0.622**	**45.30**	**0.698**	**0.8**	**49.13**
Role of the MoE						
MLP$_{\text{m}}$	0.425	0.567	42.01	0.724	0.783	46.79
ATTN$_{\text{m}}$	0.415	0.589	41.36	0.746	0.775	46.06
CNN$_{\text{m}}$	0.420	0.578	43.11	0.711	0.789	46.21
Role of Contrastive Learning						
w/o CL	0.436	0.546	42.89	0.713	0.796	47.23
Role of Fusion						
SUM	0.412	0.603	43.98	0.727	0.786	47.81
CON	0.425	0.585	41.79	0.717	0.792	47.52
ATTN	0.416	0.586	42.89	0.707	0.798	47.52
MUL	0.427	0.567	42.23	0.706	0.798	46.36

Table 3: Result of ablation experiments for different components of KHaR on CH-SIMS and MOSI datasets. V=vedio, T=text, A=audio. The best result is in bold.

4.4.1 Role of Modalities

We evaluate the contribution of each modality by comparing different combinations: bi-modal (V+T, V+A, A+T) versus tri-modal (V+A+T) inputs. As shown in Table 3, the tri-modal setting consistently achieves superior performance across all metrics, confirming the complementary nature of visual, acoustic, and textual information in affective understanding. In contrast, bi-modal configurations yield suboptimal results, with V+A particularly underperforming on MOSI, suggesting the insufficiency of non-verbal cues in capturing nuanced sentiment without textual grounding.

4.4.2 Role of the MoE

In order to verify the effectiveness of the MoE layer, we use different feature extraction methods to replace the MoE layer for comparison. As shown in Tabel 3, MLP$_{\text{m}}$ denotes a multilayer perceptron, ATTN$_{\text{m}}$ adopts a standard attention mechanism, and CNN$_{\text{m}}$ introduces convolutional operations for feature extraction. Experimental results show that replacing MoE layer with these methods leads to performance degradation, which indicates that dynamic routing and modal-specific experts of MoE layer are essential for capturing complex patterns and handling different data distributions efficiently.

4.4.3 Role of Contrastive Learning

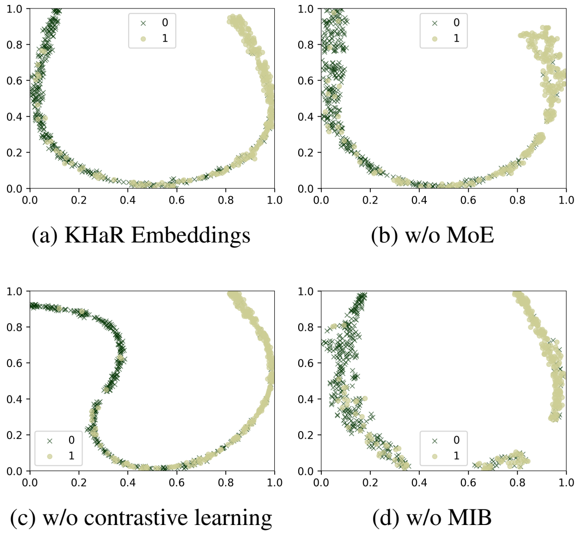
To evaluate the effectiveness of contrastive learning component, we present an ablation analysis in Table 3, where “w/o CL” denotes the removal of contrastive loss from the training objective. When removing the contrastive learning module, we observe consistent performance degradation across both SIMS and MOSI datasets, with the most pronounced drops occurring in the 5-class classification task on SIMS and the 7-class classification task on MOSI. This significant decrease demonstrates that the contrastive learning component effectively captures subtle relationships between sentiment intensities, enabling the model to better distinguish closely-related sentiment categories. The results suggest that explicitly modeling intra-modal and inter-modal feature distributions through contrastive learning is essential for maintaining discriminative power in complex sentiment identify scenarios, where traditional comtrastive learning approaches often fail to preserve these nuanced emotional relationships.

4.4.4 Role of Fusion

We compare MIB-based fusion strategy against several alternative fusion mechanisms. Among them, SUM represents the addition of different modal representations, CON represents the concatenation of different modal representations, ATTN represents the fusion using the attention mechanism, and MUL represents the multiplication of the representations of different modalities. The MIB fusion consistently achieves superior performance across most metrics, particularly on CH-SIMS. These results demonstrate that the information bottleneck encourages the extraction of task-relevant and compact representations by discarding redundant or noisy modality-specific features. This selective fusion mechanism proves more effective than naive aggregation strategies.

4.4.5 Visualizating Representations

In this section, we utilize t-SNE(Van der Maaten and Hinton, 2008) for a more intuitive visual presentation of the multimodal representation, as shown in Figure 4. Figure 4a shows the representation generated by KHaR, Figure 4b shows the representation using the MLP layers to replace the MoE layer, Figure 4c shows the representation without contrastive learning, and Figure 4d shows the representation without MIB. When the MLP layer is used to replace the MoE layer, it can be obviously seen that the representation distribution of similar samples is loose, indicating that the MoE layer can better capture the specific information within the modality and enhance the discrimination of the representation. The sample representation distribution without contrastive learning is mixed, the overlap area is large, and the lack of clear boundary, which indicates that contrastive learning makes different categories easier to distinguish after dimension reduction by bringing similar samples closer and pushing away heterogeneous samples. Samples without MIB have a more chaotic representation distribution with local small clusters, indicating poor feature consistency, and the model may overfit noise or local patterns. It shows that MIB can better filter noise information while retaining the most relevant information, and improve the generalization ability and robustness of the model.



(a) KHaR Embeddings (b) w/o MoE

(c) w/o contrastive learning (d) w/o MIB

Figure 4: Visualization of multimodal representations on CMU-MOSI. Where '0' and '1' represent positive and negative sentiment respectively.

5 Conclusion

In this paper, we propose a novel multimodal sentiment analysis framework KHaR, which is the first to extract the fine-grained information within a modality by using two steps of knowledge harvesting and refinement. This design helps the model to better capture the detailed information of a specific modality before fusion. Moreover, we introduce a dynamic contrastive mechanism based on sentiment intensity, enabling fine-grained and semantically aligned cross-modal representations. Extensive experiments across four benchmark datasets show that KHaR consistently outperforms state-of-the-art methods, achieving superior results in both accuracy and robustness. Ablation studies further highlight the indispensable role of each component. These findings position KHaR as one of the most promising and effective solutions in the field.

Limitations

Although KHaR achieves strong performance on multiple datasets, it still has some limitations. Firstly, we can consider using domain adapters and other external domain knowledge to further inject more relevant knowledge into the model. Secondly, we mainly strengthen the model's learning of fine knowledge. In the future, we can consider how to better integrate coarse-grained knowledge and fine knowledge for multimodal sentiment analysis.

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, 34:24206–24221.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. 2024. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. *arXiv preprint arXiv:2410.04491*.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 6–15.

Wei Han, Hui Chen, and Soujanya Poria. 2021b. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Jiehui Huang, Jun Zhou, Zhenchao Tang, Jiaying Lin, and Calvin Yu-Chian Chen. 2024. Tmbl: Transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowledge-Based Systems*, 285:111346.

Onno Kampman, Elham J Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction. *arXiv preprint arXiv:1805.00705*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Han Lin, Pinglu Zhang, Jiading Ling, Zhenguo Yang, Lap Kei Lee, and Wenyin Liu. 2023. Ps-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis. *Information Processing & Management*, 60(2):103229.

Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th international conference on computational linguistics*, volume 29, pages 7124–7135. Association for Computational Linguistics.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Huaishao Luo, Lei Ji, Yanyong Huang, Bin Wang, Shenggong Ji, and Tianrui Li. 2021. Scalevlad: Improving multimodal sentiment analysis via multi-scale fusion of locally descriptors. *arXiv preprint arXiv:2112.01368*.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 164–172.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-sena: An integrated platform for multimodal sentiment analysis. *arXiv preprint arXiv:2203.12441*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE.

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

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3722–3729.

Kaiwei Sun and Mi Tian. 2025. Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 40–49.

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 conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921.

Chung-Hsien Wu and Wei-Bin Liang. 2010. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21.

Yang Yang, Xunde Dong, and Yupeng Qiang. 2024. Clgsi: a multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2099–2110.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

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.

Yakun Yu, Mingjun Zhao, Shi-ang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. 2023. Conki: Contrastive knowledge injection for multimodal sentiment analysis. *arXiv preprint arXiv:2306.15796*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

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.

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)*, pages 2236–2246.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. 2023. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36:17882–17903.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*.

Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, Ken Zheng, and Qunyan Zhou. 2023. Acformer: An aligned and compact transformer for multimodal sentiment analysis. In *Proceedings of the 31st ACM international conference on multimedia*, pages 833–842.

A Datasets

CMU-MOSI is a multimodal dataset, which includes 3 modalities: text, visual and acoustic. The data was collected from 93 movie review videos on Youtube. The videos were edited into 2199 segments. Each segment is annotated with sentiment intensity in the range [-3,3]. CMU-MOSEI is similar to CMU-MOSI, but on a larger scale. It contains 23,453 annotated video clips from online video sites covering 250 different topics and 1000 different speakers. Samples in CMU-MOSEI are also labeled with sentiment intensity in the range [-3,3]. The SIMS dataset is a Chinese multimodal sentiment analysis dataset, which provides detailed annotations for each modality. The dataset includes 2,281 selected video clips from a variety of movies, TV series, and variety shows, and each sample is assigned a sentiment score ranging from -1 (extremely negative) to 1 (extremely positive). The CH-SIMS v2.0 dataset is an extension and enhancement of the CH-SIMS. The dataset collects 4402 labeled supervised data and 10161 unlabeled raw video clips from 11 different scenes. The sentiment intensity of each sample is between -1 and 1. The partition of the above dataset is shown in Tabel 4.

Dataset	#Train	#Valid	#Test	#Total	Language
CH-SIMS	1368	456	457	2281	Chinese
CH-SIMSv2	2722	647	1034	4403	Chinese
MOSI	1284	229	686	2199	English
MOSEI	16326	1871	4659	22856	English

Table 4: The statistics of CH-SIMS, CH-SIMSv2, MOSI and MOSEI.

B Experimental Setup

Here, we will mainly present the specific implementation of our experimental setup. All experiments were conducted on high performance computing nodes equipped with NVIDIA RTX 4090D GPU. On the Chinese datasets SIMS and SIMSv2, we adopt bert-base-chinese (12-layer, 768-hidden-dim) to initialize the model, while on the English datasets MOSI and MOSEI, we adopt bert-base-uncased as the baseline architecture. Both models are optimized using AdamW, including a linear warmup schedule and weight decay regularization. The main hyperparameters are shown in Table 5.

Descriptions	CH-SIMS	CH-SIMSv2	MOSI	MOSEI
Epochs	70	70	70	70
Learning Rate	3e-5	3e-5	3e-5	1e-5
Batch Size	64	64	64	36
Num of Experts	3	3	3	3
λ	1	1	0.05	0.45
Modal Dimension d_m	128	128	128	256
Fine-grained Dimension	50	50	50	50
Optimizer	AdamW	AdamW	AdamW	AdamW

Table 5: Hyperparameter Settings for different datasets.

C Derivation of MIB

The pipeline of the MIB is in Algorithm 1, and details of its formula derivation are shown as follows.

```
Algorithm 1 Multimodal Information Bottleneck
Input: Unimodal representations
$\boldsymbol{X}_m,\ m \in \{v, t, a\}$, hyper-parameter $\beta$
Output: Prediction $\hat{y}_i$
    Initialize unimodal networks $F^m$ and fusion network $F^f$;
    while not done do
        Sample a batch of utterances
        for each utterance $i$ do
            for each $m$ ($m \in \{v, t, a\}$) do
                $\boldsymbol{x}_i^m = F^m(\boldsymbol{X}_i^m; \theta_m)$
            end for
            $\boldsymbol{x}_i = F^f(\boldsymbol{x}_i^l, \boldsymbol{x}_i^a, \boldsymbol{x}_i^v; \theta_f)$
            $\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} = \mu(\boldsymbol{x}_i; \theta_\mu), \Sigma(\boldsymbol{x}_i; \theta_\Sigma)$
            $\boldsymbol{z}_i = \boldsymbol{\mu}_{z_i} + \boldsymbol{\Sigma}_{z_i} \times \boldsymbol{\epsilon}$
            $\hat{y}_i = D(\boldsymbol{z}_i; \theta_d)$
        end for
        Compute $J_{MIB}$ as in Eq. 11
    end while
```

We design the encoder $p(z \mid x)$ to be a Gaussian distribution whose mean and covariance are parameterized by a neural network:

$$p(z \mid x) = \mathcal{N}\Big(\boldsymbol{\mu}(x; \theta_\mu), \boldsymbol{\Sigma}(x; \theta_\Sigma)\Big) = \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \tag{23}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, parameterized respectively by θ_μ and θ_Σ, are neural networks designed to estimate the mean vector $\boldsymbol{\mu}_z$ and covariance matrix $\boldsymbol{\Sigma}_z$ of the Gaussian latent distribution.

Since directly sampling random variables is not conducive to gradient propagation, we use the reparameterization technique to transform the sampling process into z:

$$z = \boldsymbol{\mu}_z + \boldsymbol{\Sigma}_z \times \boldsymbol{\epsilon} \tag{24}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ denotes a sample from the standard multivariate normal distribution, and I represents an identity matrix with all diagonal elements equal to 1.

This treatment transfers the randomness to $\boldsymbol{\epsilon}$, allowing $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ to be explicitly optimized via gradients. Note that here we assume that each element in the vector z is independent from each other.For our task, we formulate $q(y \mid z)$ as:

$$\begin{aligned} q(y \mid z) &= e^{-||y-\boldsymbol{D}(z;\theta_d)||_1+C} \\ \log q(y \mid z) &= -||y-\boldsymbol{D}(z;\theta_d)||_1 + C \\ &= -||y-\hat{y}||_1 + C \end{aligned} \tag{25}$$

where $\boldsymbol{D}$ denotes a decoder parameterized by θ_d, and $\hat{y}$ is the model prediction. Here, maximizing $\log q(y \mid z)$ is equivalent to minimizing the mean absolute error (MAE) between the predicted $\hat{y}$ and the ground truth y.

In practice, MAE is frequently used to maximize the MI between the target and the latent representation z and the approximated marginal distribution of the multimodal representation z is often assumed to be a standard Gaussian distribution:

$$q(z) \sim \mathcal{N}(0, I) \tag{26}$$

Through combining Eq.23 and Eq.26, the KL divergence term $KL\big(p(z|x)||q(z)\big)$ can be evaluated as follows:

$$\begin{aligned} KL\Big(p(z|x)||q(z)\Big) &= KL\Big(\mathcal{N}(\boldsymbol{\mu}(x;\theta_\mu), \boldsymbol{\Sigma}(x;\theta_\Sigma))||\mathcal{N}(0,I)\Big) \\ &= KL\Big(\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)||\mathcal{N}(0,I)\Big) \end{aligned} \tag{27}$$

where this formulation relies on the assumption that the reparameterizations of both $p(z \mid x)$ and $q(z)$ are chosen such that the KL divergence admits a closed-form analytical expression.

To approximate the integral over x, z and y, we employ Monte Carlo sampling, which allows the overall objective J_{MIB} (Eq. 11) to be rewritten in the following simplified form:

$$J_{\text{MIB}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}\mathsf{L}_i - \beta \cdot \text{KL}\left(\mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \,\|\, \mathcal{N}(0,I)\right)\right] \tag{28}$$

where n refers to the number of samples (i.e., the batch size), and the index i denotes the individual data point in the sampled batch.

Maximizing this objective can maximize the discrimination ability of the target variable and effectively compress the redundant information in the multimodal representation x.