# Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis

**Xinyu Feng   Yuming Lin*   Lihua He   You Li   Liang Chang   Ya Zhou**
Guangxi Key Laboratory of Trusted Software,
Guilin University of Electronic Technology, Guangxi, China

## Abstract

Multimodal Sentiment Analysis (MSA) utilizes multimodal data to infer the users' sentiment. Previous methods focus on equally treating the contribution of each modality or statically using text as the dominant modality to conduct interaction, which neglects the situation where each modality may become dominant. In this paper, we propose a Knowledge-Guided Dynamic Modality Attention Fusion Framework (KuDA) for multimodal sentiment analysis. KuDA uses sentiment knowledge to guide the model dynamically selecting the dominant modality and adjusting the contributions of each modality. In addition, with the obtained multimodal representation, the model can further highlight the contribution of dominant modality through the correlation evaluation loss. Extensive experiments on four MSA benchmark datasets indicate that KuDA achieves state-of-the-art performance and is able to adapt to different scenarios of dominant modality.[1]

## 1 Introduction

Since users' sentiment expressions are reflected in multiple modalities on social media, multimodal sentiment analysis (MSA) has garnered rising attention in recent years. It aims to mine and comprehend the sentiments of online videos (Soleymani et al., 2017; Zeng et al., 2021; Kaur and Kautish, 2022). Most recent MSA methods can be grouped into two categories: ternary symmetric-based methods (Zadeh et al., 2017; Liu et al., 2018; Zadeh et al., 2018a; Tsai et al., 2019; Hazarika et al., 2020; Yu et al., 2021; Sun et al., 2022; Huang et al., 2024) and text center-based methods (Han et al., 2021a,b; Wang et al., 2022; Li et al., 2022; Lin and Hu, 2022; Wang et al., 2023; Zhang et al., 2023). Ternary symmetric-based methods focus on equally
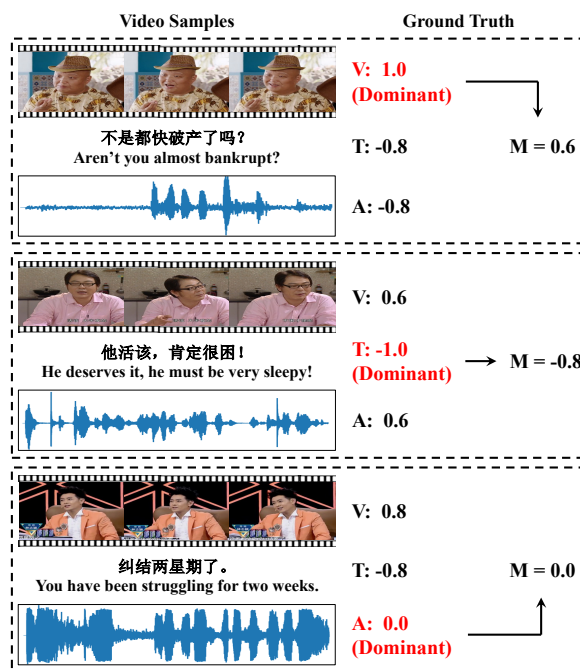
Figure 1: Three video samples with vision, text, or audio as the dominant modality from CH-SIMS dataset. In the Ground Truth, V, T and A denote the vision, text and audio, respectively. M is the overall sentiment label of the video sample. The values of all labels are from -1 (negative) to 1 (positive).

treating the contribution of each modality and modeling the bidirectional relationship of all modality pairs. Text center-based methods focus on using text as the dominant modality to guide the vision and audio modalities to interact with it to adjust the contributions of different modalities properly. Thus, both ternary symmetric-based methods and text center-based methods consider the distribution of importance among modalities to be static and fix the dominant modality.

However, as shown in Figure 1, we discovered that in certain situations, vision, text, or audio could be the dominant modality respectively. With the first sample in Figure 1, since the vision label is more consistent with the overall sentiment label,

it is dominant. Based on an investigation of commonly used datasets in MSA, we find that these situations are not uncommon. Detailed statistics and analysis can be found in Appendix A. Thus, when the dominant modality is not fixed, the ternary symmetric-based methods cannot effectively adapt to the situation where any modality is dominant because they do not consider the differences of importance between modalities. The text center-based methods statically set text as the dominant modality, and when other modalities are dominant, the model's attention is distracted by the text.

In this paper, we propose a **K**nowledge-G**u**ided **D**ynamic Modality **A**ttention Fusion Framework (KuDA), which improves the model performance and makes it adaptable to more complex and wider scenarios by dynamically selecting the dominant modality and adjusting the contributions of each modality according to different samples. Specifically, KuDA first uses the BERT model and two Transformer Encoders to extract semantic features of text, vision, and audio modalities. Then, KuDA performs sentiment knowledge injection and sentiment ratio conversion by the adapters and decoders, which can extract sentiment clues and guide KuDA in selecting the dominant modality further. Next, the dynamic attention fusion module is designed to capture similar sentiment information and gradually adjusts the attention weights between modalities by interacting sentiment knowledge with different levels of multimodal features. Based on the correlation evaluation between the multimodal features and the unimodal features, we use Noise-Contrastive Estimation (Gutmann and Hyvärinen, 2010) to highlight the contribution of the dominant modality further. Finally, KuDA predicts the sentiment score through a multilayer perceptron.

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

- We propose KuDA, a Knowledge-Guided Dynamic Modality Attention Fusion Framework for multimodal sentiment analysis, which improves the performance by dynamically selecting the dominant modality, making model adaptable to complex and wide scenarios.

- We design a Dynamic Attention Fusion module, which utilizes sentiment knowledge to guide different levels of multimodal features and achieves dynamic fusion by adjusting the contribution of each modality.

- Extensive experiments on four MSA datasets show that KuDA achieves state-of-the-art performance. We further analyze the experimental results to prove the effectiveness of KuDA.

## 2 Related Work

In this section, we briefly overview related work in ternary symmetric-based methods and text center-based methods.

**Ternary symmetric-based methods.** Previous works in this category have primarily focused on modeling bidirectional relationships in each modality pair and treating the contribution of each modality equally. For instance, some researchers designed the fusion architectures of tensor-based (Zadeh et al., 2017; Liu et al., 2018), LSTMs-based (Zadeh et al., 2018a) and mlp-based (Sun et al., 2022) to capture the commonalities between modalities. While Tsai et al. (2019) designed MuLT, which is a transformer-based model, to find similar information between each modality pair. Some studies (Tsai et al., 2018; Hazarika et al., 2020; Yang et al., 2022) project each modality into two subspaces, separately learning the commonalities and characteristics of the modalities to aid the fusion process. Multi-label strategy (Yu et al., 2021) and contrastive learning (Mai et al., 2023) are introduced to improve the quality of unimodal features. In addition, Huang et al. (2024) designed TMBL, which can handle bimodal and trimodal features to capture and utilize bound modal features.

**Text center-based methods.** Previous works in this category mainly focus on improving the quality of fuse representation through the text modality to guide the vision and audio modalities. For example, Delbrouck et al. (2020), Han et al. (2021a) and Wang et al. (2023) use the transformer-based method to integrate similar information from other modalities by text modality. While some researchers (Rahman et al., 2020; Wang et al., 2022) enhance the text representations by integrating vision and audio information into a pretrained language model. Moreover, to decrease the potential sentiment-irrelevant and conflicting information, some studies reduce additional noise by maximizing mutual information (Han et al., 2021b) or adaptive representation learning (Zhang et al., 2023). Meanwhile, contrastive learning is introduced to learn invariant and similar information between text with other modalities (Li et al., 2022; Lin and Hu, 2022).
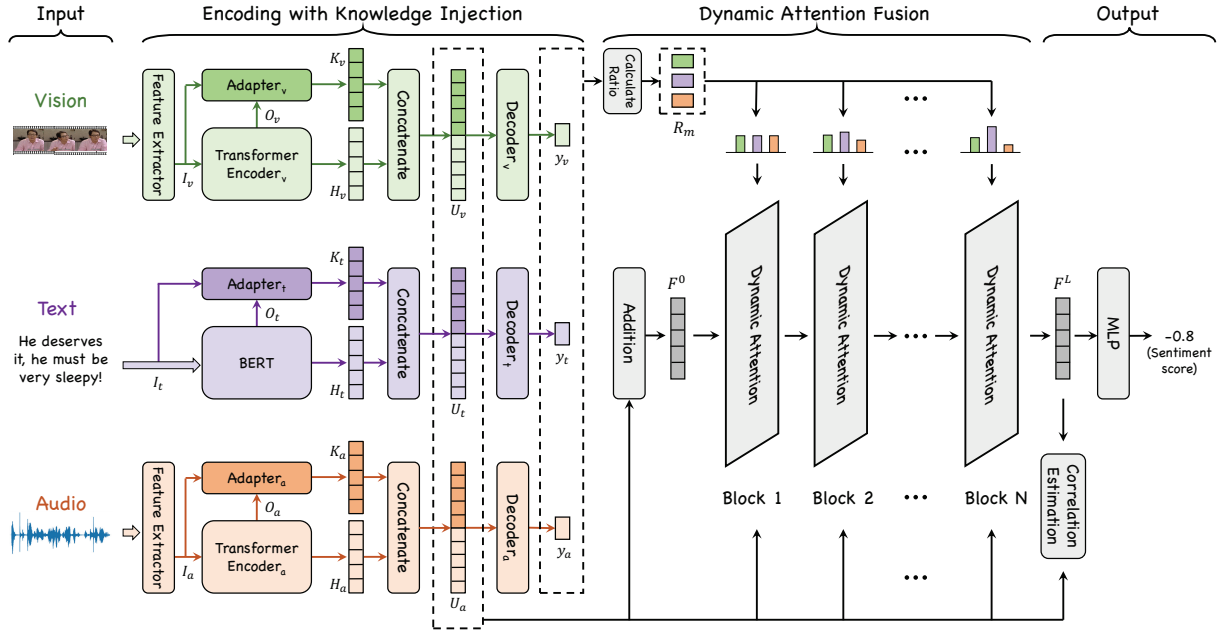
Figure 2: The overall architecture of KuDA. The green, purple, orange, and gray modules represent the relevant operations of vision, text, audio and multimodal fusion.

Despite the promising results achieved for MSA, the above approaches treat each modality in a balanced way or statically set text as the dominant modality. This causes the approach to be distracted by the secondary modalities, hindering it from dynamically adjusting to diverse scenarios where the dominant modality varies, which limits the performance of MSA.

## 3 Methodology

### 3.1 Overall Architecture

As shown in Figure 2, which shows the overall workflow of KuDA. Specifically, KuDA first extracts unimodal low-level features from the raw multimodal input. Then, the adapters and encoders extract unimodal high-level features and learn sentiment knowledge simultaneously. We utilize the decoders to predict the unimodal sentiments and convert them into sentiment ratios to guide dynamic fusion. Next, we designed a dynamic attention fusion module, which selects the dominant modality according to different scenarios and dynamically adjusts attention weights with sentiment ratios and knowledge representations. Finally, the multimodal representation is used to conduct the MSA task by multilayer perceptron (MLP) and estimate correlation with knowledge representations.

In addition, to guide the model in adjusting the attention weights by sentiment knowledge, KuDA adopts a two-stage training method.

### 3.2 Problem Definition and Notations

In MSA task, the input data consists of text $(t)$, vision $(v)$ and audio $(a)$ modalities. The sequences of three modalities are represented as triplet $(I_t, I_v, I_a)$, which include $I_t \in \mathbb{R}^{T_t \times d_t}$, $I_v \in \mathbb{R}^{T_v \times d_v}$, and $I_a \in \mathbb{R}^{T_a \times d_a}$, where $T_m, m \in \{t, v, a\}$ is the sequence length and $d_m$ represents the vector dimension. The prediction is the sentiment score $\hat{y}$, which is a discrete value between [-1, 1] and [-3, 3], with values greater than, equal to, and less than 0 representing positive, neutral, and negative, respectively.

### 3.3 Encoding with Knowledge Injection

We encode the input of each modality $I_{m \in \{t,v,a\}}$ into the global semantic representations $H_m \in \mathbb{R}^{T_m \times d_m}$ and the knowledge-sentiment representations $K_m \in \mathbb{R}^{T_m \times d_m}$ via the pretrained encoders and adapters, respectively.

**Global semantic representations.** For the text modality, to effectively extract from low-level to high-level text semantic information and facilitate subsequent knowledge inject with Adapter, we use the BERT (Kenton and Toutanova, 2019) to encode the input sentences $I_t$, and extract the hidden state of the last layer as the global semantic representation $H_t$:
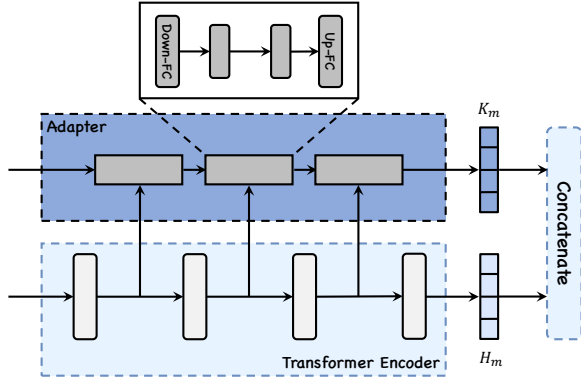
Figure 3: The architecture of the BERT, transformer encoder, adapter, and their connection. The vertical gray rectangles denote the Transformer Encoder Layer. Down-FC and Up-FC represent the fully connected layers (FC) used to decrease and increase the dimension.

$$H_t, O_t = \text{BERT}\left(I_t; \theta^{\text{BERT}}\right) \tag{1}$$

where $O_t$ denotes the hidden states except the last layers. Similar to text modality, for the vision and audio modalities, we use stacked transformer encoder layers (Vaswani et al., 2017) to capture global semantic representations $H_m, m \in \{v, a\}$:

$$H_m, O_m = \text{Encoder}_m\left(I_m; \theta_m^{\text{encoder}}\right) \tag{2}$$

Since the obtained $O_m, m \in \{t, v, a\}$ mainly include the general knowledge, they are input into the Adapter to inject the sentiment information.

**Knowledge-sentiment representations.** Since the adapter is commonly used to enhance the knowledge cognition of pretrained language models (Wei et al., 2021), we use it to inject unimodal sentiment knowledge. In KuDA, the adapter is plugged outside of the encoder and stacked with identical blocks, and its detailed architecture is shown in Figure 3. We connect each transformer encoder layer of the vision and audio modalities to the Adapter. To prevent information redundancy, we select part of the intermediate layers of BERT to connect. For the first adapter block, we take the unimodal features $I_{m \in \{t, v, a\}}$ as one of the inputs. The output of adapter is denoted as knowledge-sentiment representation $K_m, m \in \{t, v, a\}$:

$$K_m = \text{Adapter}_m\left(I_m, O_m; \theta_m^{\text{adapter}}\right) \in \mathbb{R}^{T_m \times d_m} \tag{3}$$

where $\theta_m^{\text{Adapter}}$ is denoted the pretrained parameters of the adapter of $m$ modality.

**Unimodal sentiment scores.** We combine the above two representations to obtain the knowledge-enhanced representation of each modality $U_m, m \in \{t, v, a\}$. Then, we use this representation to predict the unimodal sentiment score $\hat{y}_m$ by a decoder consisting of MLP:

$$U_m = [K_m; H_m] \in \mathbb{R}^{T_m \times 2d_m} \tag{4}$$

$$\hat{y}_m = \text{Decoder}_m\left(U_m; \theta_m^{\text{decoder}}\right) \tag{5}$$

where $[\cdot; \cdot]$ denotes the concatenation. Due to the difference between the unimodal and multimodal sentiment scores can indicate the effective amount of information provided by the corresponding modality, so we use the sentiment scores $\hat{y}_m$ to guide attention weights further.

### 3.4 Dynamic Attention Fusion

#### 3.4.1 Unimodal Sentiment Ratio

Since the difference in sentiment score of unimodal $y_m$ and multimodal $y$ is inversely proportional to the weight of unimodal, we choose the inverse proportional function $\exp(-kx)$ and normalization operation, and utilize the ground truth of MSA $y$ to convert the unimodal sentiment score $\hat{y}_m$ into sentiment ratio $R_m, m \in \{t, v, a\}$ during training to guide subsequent dynamic fusion:

$$\begin{aligned} D_m &= \exp\left(-k\left|\hat{y}_m - y\right|^2\right) \\ R_m &= \frac{D_m}{D_t + D_v + D_a} \end{aligned} \tag{6}$$

where $k$ denotes the function's slope, which can scale the sentiment ratio. Due to the model effectively learned how to adjust the contribution between modalities during training, we fixed the sentiment ratio to 1 during the test stage to highlight the model's ability to adjust weights.

#### 3.4.2 Dynamic Attention Block

In order to unify the length and dimension axis of the unimodal knowledge-enhanced representation $U_m \in \mathbb{R}^{T_m \times 2d_m}$ for multimodal fusion, we utilize three projectors to obtain the updated knowledge-enhanced representation of each modality $\overline{U}_m, m \in \{t, v, a\}$. In addition, due to any one or more of the text, vision, and audio modalities may become the dominant modality, we first sum the obtained representations $\overline{U}_m$ as the input of the first dynamic attention block $F^0$:
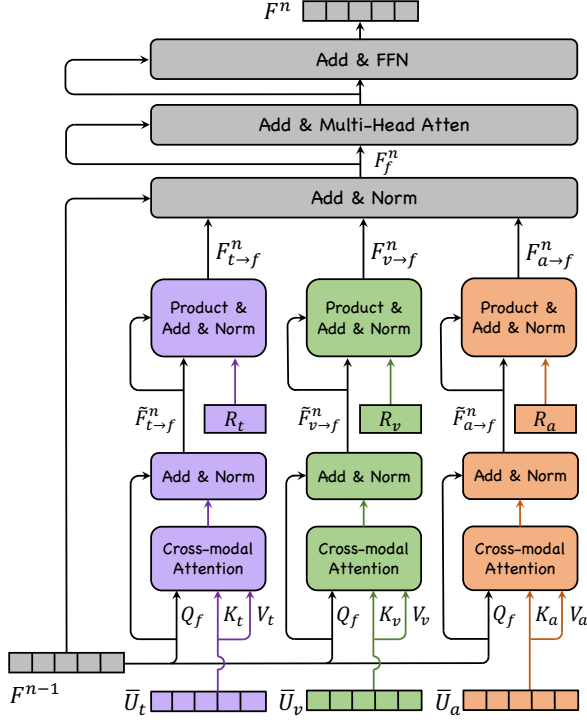
Figure 4: Architecture of a single dynamic attention block. The purple, green, and orange represent the interaction of multimodal representation with text, vision, and audio modalities, respectively.

$$\overline{U}_m = \text{Projector}_m\left(U_m\right) \in \mathbb{R}^{T_f \times d_f}$$
$$F^0 = \overline{U}_t + \overline{U}_v + \overline{U}_a \in \mathbb{R}^{T_f \times d_f} \quad (7)$$

where $\text{Projector}_m$ consists of two linear layers, $T_f$ and $d_f$ denote length and dimension in fusion stage.

We then stack the dynamic attention blocks to form a pipeline, which is shown in Figure 2. At the same time, we use the output of the prior block $F^{n-1}$, the knowledge-enhanced representation $\overline{U}_m$, and the sentiment ratio $R_m, m \in \{t, v, a\}$ as the input of the next block and obtain its output $F^n$:

$$F^n = \text{DAB}\left(F^{n-1}, \overline{U}_m, R_m\right) \in \mathbb{R}^{T_f \times d_f} \quad (8)$$

where DAB is the dynamic attention block. This is because the representations $\overline{U}_m$ have richer sentiment clues, and the sentiment ratios $R_m$ can indicate the contribution of each modality. Finally, we take the output of the last block as the final multimodal representation $F^L \in \mathbb{R}^{T_f \times d_f}$ and use it to conduct the MSA task.

As shown in Figure 4, to adjust the weights of each modality, we designed a dynamic attention block. Specifically, we first introduce the cross-modal attention (CAttn), which gradually determines the dominant modality by capturing the

amount of similar information between unimodal representations $\overline{U}_m, m \in \{t, v, a\}$ and multimodal representation $F^{n-1}$. Since $Q$ of attention is used to specify the location of attention, we take the multimodal features as $Q$, and the unimodal features as $K$ and $V$, and perform the Layer Norm (LN):

$$\tilde{F}_{m \to f}^n = \text{LN}\left(F^{n-1} + \text{CAttn}\left(F^{n-1}, \overline{U}_m, \overline{U}_m\right)\right) \quad (9)$$

Next, since the sentiment ratio $R_m$ can further guide the dynamic fusion, we multiply it with the middle representation $\tilde{F}_{m \to f}^n, m \in \{t, v, a\}$. We then sum the obtained representations and the multimodal representation of input $F^{n-1}$ to fine-tune the contributions of different modalities:

$$F_{m \to f}^n = \text{LN}\left(\tilde{F}_{m \to f}^n + \left(R_m \times \tilde{F}_{m \to f}^n\right)\right) \quad (10)$$

$$F_f^n = F^{n-1} + \text{LN}\left(\sum_{m \in \{t,v,a\}} F_{m \to f}^n\right) \quad (11)$$

At last, we input the $F_f^n$ into the multi-head attention and feedforward neural network to get the output of dynamic attention block $F^n \in \mathbb{R}^{T_f \times d_f}$.

### 3.5 Output and Training Objectives

To further improve the utilization of the dominant modality, we estimate the correlation of the multimodal representation $F^L$ and the unimodal representations $\overline{U}_m, m \in \{t, v, a\}$ through the Contrastive Predictive Coding (Oord et al., 2018), and integrated it into the Noise-Contrastive Estimation framework (Gutmann and Hyvärinen, 2010) to form the correlation estimation (CE) loss $\mathcal{L}_{cor}$:

$$\mathcal{L}_{cor} = \sum_{m \in \{t,v,a\}} \mathcal{L}_{\text{NCE}}\left(F^L, \overline{U}_m\right) \quad (12)$$

In the Output, we input the representation $F^L$ into a MLP to predict sentiment score $\hat{y}$. Given the predictions $\hat{y}$ and the ground truth $y$, we calculate the MSA task loss $\mathcal{L}_{reg}$ by mean absolute error. Finally, we training KuDA by the union loss $\mathcal{L}_{task}$:

$$\hat{y} = \text{MLP}\left(\text{Mean}\left(F^L\right)\right) \quad (13)$$

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (14)$$

$$\mathcal{L}_{task} = \mathcal{L}_{reg} + \alpha\mathcal{L}_{cor} \quad (15)$$

| 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 1: The statistics of CH-SIMS, CH-SIMSv2, MOSI and MOSEI.

| Descriptions | CH-SIMS | CH-SIMSv2 | MOSI | MOSEI |
|--------------|---------|-----------|------|-------|
| Batch Size | 32 | 32 | 32 | 64 |
| Initial Learning Rate | 3e-5 | 3e-5 | 3e-5 | 4e-5 |
| Knowledge Injection of Bert | 3,6,9,11 | 3,6,9,11 | 6,9 | 6,9 |
| Dynamic Transformer Block | 3 | 3 | 2 | 4 |
| Vector Dimension $F^i$ | 256 | 256 | 256 | 256 |
| k | 0.1 | 0.3 | 2.0 | 5.0 |
| $\alpha$ | 0.01 | 0.01 | 0.01 | 0.1 |
| Epochs | 50 | 50 | 50 | 50 |
| Optimizer | AdamW | AdamW | AdamW | AdamW |

Table 2: Hyper-parameters settings on different datasets.

where Mean $(\cdot)$ denotes the average operation in length axis. $\alpha$ is a parameter that balances the contribution of different losses. The specific training algorithm of KuDA can be found in Appendix B.

## 4 Experiments

### 4.1 Datasets and Evaluation Metrics

We conduct experiments on four publicly benchmark datasets of MSA, including CH-SIMS (Yu et al., 2020), CH-SIMSv2 (Liu et al., 2022), MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018b). The statistic details of four datasets are shown in Table 1.

Following previous works (Hazarika et al., 2020; Yu et al., 2021; Zhang et al., 2023), we used the accuracy of 3-class (Acc-3) and 5-class (Acc-5) on CH-SIMS and CH-SIMSv2, the accuracy of 7-class (Acc-7) on MOSI and MOSEI, and the accuracy of 2-class (Acc-2), Mean Absolute Error (MAE), Pearson Correlation (Corr), and F1-score (F1) on all datasets. Moreover, on MOSI and MOSEI, Acc-2 and F1 used two calculation ways: negative/non-negative (has-0) and negative/positive (non-0). Except for MAE, higher values indicate better performance for all metrics.

### 4.2 Baselines

To validate the KuDA's performance, we conduct a fair comparison with several competitive and state-of-the-art (SOTA) baselines, including the ternary symmetric-based methods: TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MuLT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), CubeMLP (Sun et al., 2022) and TMBL (Huang et al., 2024), and the text center-based methods: MMIM (Han et al., 2021b), BBFN (Han et al., 2021a), CENet (Wang et al., 2022), TETFN (Wang et al., 2023) and ALMT (Zhang et al., 2023).

### 4.3 Experimental Settings

To ensure fairness with other baselines, we follow recent competitive and SOTA methods to set the

proposed method. The training stages consist of the unimodal stage and the multimodal stage. In addition, we adopt BERT to extract the features of text modality where "bert-base-chinese"[2] is employed for CH-SIMS and CH-SIMSv2 and "bert-base-uncased"[3] is employed for MOSI and MOSEI. In the vision and audio modalities, we directly use the features provided by the original datasets. Moreover, we develop the KuDA using a single NVIDIA RTX 3090 GPU for all datasets. The detailed setting of the best hyper-parameters can be referred to in Table 2.

### 4.4 Performance Comparison

Table 3 and Table 4 present the comparison results of the baselines and the proposed method on CH-SIMS, CH-SIMSv2, MOSI and MOSEI.

Since the distribution of modality importance in the CH-SIMS and CH-SIMSv2 is more uniform, they are more complex than MOSI and MOSEI. As shown in Table 3, the proposed method outperforms all baselines on all metrics. It is worth noting that our method achieves superior performance on the CH-SIMSv2 dataset. For example, compared with ALMT (text center) and TMBL (ternary symmetric), our method achieves 8.32% and 9.19% improvement on the Acc-5 and also achieves significant improvement on the Acc-3. Therefore, achieving superior performance in the more challenging scenario indicates that KuDA can adjust the distribution of modality weights to complete dynamic fusion. It also shows that adjusting the dominant modality is crucial for MSA. Furthermore, although the importance of modalities is not evenly distributed in MOSI and MOSEI, and the text modality plays an important role, it can be seen from Table 4 that KuDA still obtained SOTA performance in almost all metrics. At the same time, KuDA surpasses some text-center training methods, such as BBFN and ALMT.

---

[2]https://huggingface.co/bert-base-chinese
[3]https://huggingface.co/bert-base-uncased

| 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 |
| **KuDA** | **0.408** | **0.613** | **43.54** | **66.52** | 80.74 | **80.71** | **0.271** | **0.759** | **61.22** | **76.21** | **82.11** | **82.04** |

Table 3: Performance comparison on CH-SIMS and CH-SIMSv2. Note: the best result is marked in bold; [†] means the result is from (Mao et al., 2022); [*] denotes the results are reproduced from code provided by their authors.

| Methods | MOSI | | | | | MOSEI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-7 | Acc-2 | F1 | MAE | Corr | Acc-7 | 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 |
| MulT[†] | 0.879 | 0.702 | 36.91 | 79.71/80.98 | 79.63/80.95 | 0.559 | 0.733 | 52.84 | 81.15/84.63 | 81.56/84.52 |
| MISA[†] | 0.776 | 0.778 | 41.37 | 81.84/83.54 | 81.82/83.58 | 0.557 | 0.751 | 52.05 | 80.67/84.67 | 81.12/84.66 |
| 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 |
| Self-MM[†] | 0.708 | **0.796** | 46.67 | 83.44/85.46 | 83.36/85.43 | 0.531 | 0.764 | **53.87** | **83.76**/85.15 | 83.82/84.90 |
| 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 |
| 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 |
| **KuDA** | **0.705** | 0.795 | **47.08** | **84.40/86.43** | **84.48/86.46** | **0.529** | **0.776** | 52.89 | 83.26/**86.46** | 82.97/**86.59** |

Table 4: Performance comparison on MOSI and MOSEI. Note: the best result is highlighted in bold; in Acc-2 and F1, the left of the / corresponds to "negative/non-negative" and the right corresponds to "negative/positive"; [†] means the result is from (Mao et al., 2022); [*] denotes the results are reproduced from code provided by their authors.

| Methods | CH-SIMSv2 | | | MOSI | | |
|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-5 | MAE | Corr | Acc-7 |
| **KuDA** | **0.271** | **0.759** | **61.22** | **0.705** | 0.795 | **47.08** |
| w/o KIP | 0.288 | 0.729 | 56.87 | 0.731 | 0.798 | 44.61 |
| w/o Adapter | 0.286 | 0.735 | 57.54 | 0.714 | 0.787 | 46.79 |
| w/o EKI | 0.293 | 0.733 | 56.48 | 0.742 | 0.778 | 44.46 |
| w/o SR | 0.281 | 0.736 | 58.12 | 0.729 | 0.798 | 45.19 |
| w/o DAF | 0.309 | 0.716 | 54.06 | 0.754 | 0.779 | 43.73 |
| w/o CE Loss | 0.277 | 0.752 | 59.38 | 0.712 | **0.799** | 44.31 |

Table 5: Ablation results of KuDA's components on CH-SIMSv2 and MOSI. Note: "KIP" is the Knowledge Inject Pretraining; "EKI" is denoted the Encoding with Knowledge Injection module; "SR" denotes the Sentiment Ratio; "DAF" denotes the Dynamic Attention Fusion module; the best result is highlighted in bold.

## 4.5 Ablation Study and Analysis

### 4.5.1 Effects of Different Components

We conducted ablation studies to validate the effectiveness of each component, as shown in Table 5. By comparing the "w/o DAF" and KuDA, we observe that removing the DAF can seriously reduce performance. This means that KuDA dynamically adjusts the attention weights between modalities for different scenarios to select the dominant modal-ity. In addition, the performance decreases in "w/o EKI", which shows that sentiment knowledge can further guide dynamic fusion. The performance decreases after removing other modules, showing their effectiveness. Since improving the utilization of vision and audio on MOSI introduces noise, Corr has improved.

### 4.5.2 Importance of Different Modalities

To validate the impact of text, vision and audio modalities, we performed ablation studies that removed each modality on KuDA, ALMT and Cube-MLP, as shown in Table 6.

In CH-SIMSv2, which has more complex scenes, we can see that KuDA can reach the SOTA when each modality is removed. Meanwhile, when performance degrades, KuDA can still achieve acceptable results. In contrast, the other baselines will have a significant performance degradation, which proves that our method can adaptive focus on the suboptimal modality to capture sentiment features. For the MOSI dataset, which is mainly text, we can see that CubeMLP has dropped significantly when each modality is removed. Furthermore, ALMT

14761

| Methods | CH-SIMSv2 | | | MOSI | | |
|---|---|---|---|---|---|---|
| | MAE | Corr | Acc-5 | MAE | Corr | Acc-7 |
| V+A | 0.362 | 0.562 | 47.20 | 1.370 | 0.235 | 17.20 |
| T+V | 0.298 | 0.721 | 54.93 | 0.769 | 0.770 | 43.29 |
| T+A | 0.335 | 0.645 | 52.13 | 0.733 | **0.795** | 42.27 |
| **KuDA** | **0.271** | **0.759** | **61.22** | **0.705** | **0.795** | **47.08** |
| V+A | 0.451 | 0.449 | 39.94 | 1.437 | 0.201 | 15.74 |
| T+V | 0.346 | 0.623 | 48.26 | 0.772 | 0.778 | 43.15 |
| T+A | 0.368 | 0.586 | 46.03 | 0.736 | 0.788 | 43.88 |
| **ALMT** | **0.308** | **0.700** | **52.90** | **0.712** | **0.793** | **46.79** |
| V+A | 0.452 | 0.361 | 38.97 | 1.453 | 0.137 | 15.45 |
| T+V | 0.354 | 0.615 | 48.94 | 0.796 | 0.745 | 41.98 |
| T+A | 0.362 | 0.623 | 47.87 | 0.816 | 0.739 | 41.54 |
| **CubeMLP** | **0.334** | **0.648** | **52.90** | **0.755** | **0.772** | **43.44** |

Table 6: Importance of different modalities on KuDA, ALMT (text center) and CubeMLP (ternary symmetric). T, V, and A represent text, vision, and audio modalities. Note: the best result is highlighted in bold.

| Fusion Methods | MAE | Corr | Acc-5 | Acc-3 | Acc-2 | F1 |
|---|---|---|---|---|---|---|
| Concatenation | 0.296 | 0.728 | 79.98 | 72.53 | 54.64 | 79.93 |
| Addition | 0.304 | 0.721 | 78.14 | 72.24 | 54.55 | 78.02 |
| Tensor Fusion (TFN) | 0.282 | 0.754 | 80.75 | 75.05 | 55.80 | 80.64 |
| Low-rank Fusion (LMF) | 0.321 | 0.711 | 79.50 | 72.24 | 53.09 | 79.48 |
| CMT (BBFN, ALMT) | 0.282 | 0.745 | 81.43 | 74.08 | 57.16 | 81.33 |
| **KuDA** | **0.271** | **0.759** | **82.11** | **76.21** | **61.22** | **82.04** |

Table 7: The performance of different fusion methods on CH-SIMSv2. Note: the CMT denotes Cross-modal Transformer.

drops sharply after removing text and is lower than KuDA. Notably, observing the performance degradation trend after removing a certain modality indicates that the importance of each modality is evenly distributed in the CH-SIMSv2, while the importance of text modality is higher in MOSI.

### 4.5.3 Effects of Different Fusion Methods

To analyze the effects of different fusion techniques, we conducted some experiments shown in Table 7. Obviously, when faced with complex scenes, using either ternary symmetric-based (TFN, LMF) or text center-based (BBFN, ALMT) fusion methods will result in performance decline. This indicates that not focusing on the dominant modality or statically setting the dominant modality will limit the performance of MSA. However, the use of our Dynamic Attention Fusion to dynamically fuse unimodal features is the most effective.

### 4.5.4 Effects of Correlation Estimation

As shown in Figure 5, we discuss the impact of CE loss on CH-SIMSv2 and MOSI by modifying the $\alpha$. We compare the MAE, Acc-5 and Acc-7 as these metrics indicate the method's ability to
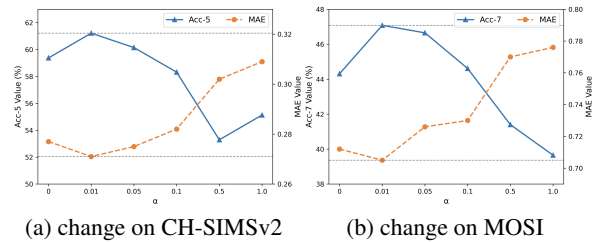


(a) change on CH-SIMSv2     (b) change on MOSI

Figure 5: Visualization of performance with change $\alpha$ on CH-SIMSv2 and MOSI.


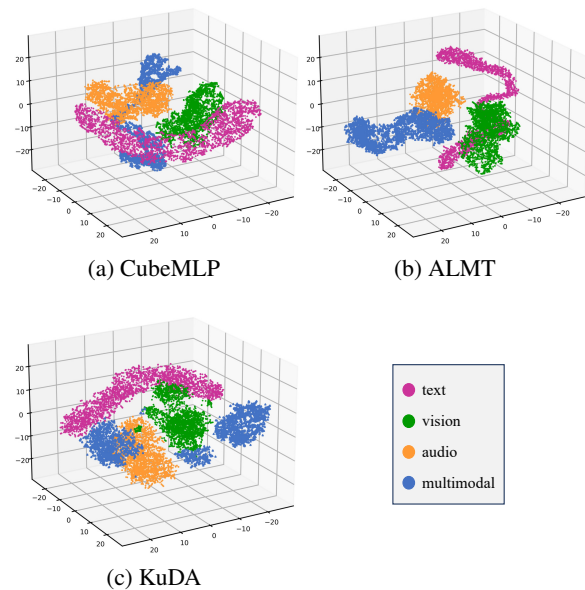
(a) CubeMLP     (b) ALMT



(c) KuDA

Figure 6: The t-SNE visualization (van der Maaten and Hinton, 2008) of the unimodal features (text, vision, and audio) and multimodal features in (a) CubeMLP; (b) ALMT; (c) KuDA.

predict fine-grained sentiment.

Compared to removing CE loss, i.e. $\alpha$=0, the model's performance achieves STOA when using CE loss and setting $\alpha$=0.01. This shows that CE loss can highlight the contribution of the dominant modality further. However, the performance shows a downward trend when $\alpha$ increases, indicating that KuDA will enhance the retention of non-dominant modality features in the multimodal representation when CE loss increases, limiting its performance.

### 4.5.5 Visualization of Features Distribution

To verify KuDA can dynamically select dominant modality, we use t-SNE to visualize the features of text, vision, audio and multimodal on CH-SIMSv2, as shown in Figure 6. We then selected two typical methods, CubeMLP (ternary symmetric) and ALMT (text center), to compare with KuDA.

| Cases | Case (a) | Case (b) | Case (c) |
|---|---|---|---|
| Vision |  |  |  |
| Text | Aren't you almost bankrupt? | He deserves it, he must be very sleepy! | You have been struggling for two weeks. |
| Audio |  |  |  |
| Ground Truth | V: 1.0　　T: -0.8　　A: -0.8<br>M = 0.6 | V: 0.6　　T: -1.0　　A: 0.6<br>M = -0.8 | V: 0.8　　T: -0.8　　A: 0.0<br>M = 0.0 |
| Output | 0.4 | -1.0 | 0.0 |
| Attention Weight |  |  |  |

Table 8: Three cases from CH-SIMS. Ground Truth consists of unimodal labels (V, T, and A) and the multimodal label (M). Output shows the multimodal prediction from KuDA. Attention Weight shows the distribution of attention weights from the multimodal features with the features of vision, text, and audio modalities. In Attention Weight, the left, middle, and right figures of each case represent the multimodal and vision, text, and audio modality weights, respectively. The partial attention of the dominant modality is marked with a dashed box.

In Figure 6a, we can see that since CubeMLP treats contributions of each modality equally, all unimodal features are averaged around the multimodal features. In addition, as shown in Figure 6b, due to ALMT is a text center-based method, we can observe that the text features is in the middle of the audio and vision, and all unimodal features are distributed on the other side of multimodal features. However, as can be seen from Figure 6c, the difference is that KuDA's multimodal features is divided into three clusters and is close to the text, audio, and vision features respectively. This indicates that KuDA dynamically selects the dominant modality to make the multimodal features closer to it.

### 4.5.6 Case Study

To better prove that the our method can dynamically adjust contributions of different modalities, we selected three challenging cases for further analysis, as shown in Table 8.

We can observe that in case (a), although the text and audio express stronger negative sentiments, KuDA can still output the correct prediction. This case shows that by adjusting vision as the dominant modality, KuDA effectively captures the speaker's information of expression and action, which also guides the fusion of text and audio modalities. In cases (b) and (c), the similar distributions of labels also occurred. KuDA still makes correct predictions, which indicates that it captures the semantic information of text in case (b) and the intonation

information of audio in case (c) by adjusting the attention weights. Meanwhile, it can be seen in the Attention Weight of Table 8 that attention weight for the dominant modality (there are denser dark blocks) is higher than that for the other modalities. This once again proves the importance of dynamic attention fusion for the MSA task.

## 5 Conclusion

In this paper, we propose a Knowledge-Guided Dynamic Modality Attention Fusion Framework (KuDA) to simultaneously solve the MSA task of the modality importance being equally or unequally distributed. Since KuDA dynamically adjusts the contribution of each modality for different scenarios, it effectively improves the utilization of the dominant modality. This enables our model to be more effective and generalized on four popular MSA benchmark datasets. At last, we perform comprehensive ablation studies to analyze this phenomenon.

## Limitations

Although the KuDA proposed in this paper has yielded exceptional outcomes, there remain several limitations that offer opportunities for further enhancement. First, KuDA suffers from an error propagation problem due to its two-stage training method. When the pretrained sentiment knowledge incorrectly predicts the sentiment score, the sentiment ratio will introduce noise when the model

adjusts the weights. Second, KuDA needs to be pretrained using the sentiment knowledge of each modality, which increases the resource consumption of model training. In future work, we will further try to explore fine-tuning the pretrained knowledge injection module in the prediction stage to solve the above limitations.

## Acknowledgements

## References

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Wei Han, Hui Chen, Alexander F. Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *ICMI '21: 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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event*, pages 9180–9192.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event*, pages 1122–1131.

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.

Ramandeep Kaur and Sandeep Kautish. 2022. Multimodal sentiment analysis: A survey and comparison. *Research anthology on implementing sentiment analysis across multiple disciplines*, pages 1846–1870.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ziming Li, Yan Zhou, Weibo Zhang, Yaxin Liu, Chuanpeng Yang, Zheng Lian, and Songlin Hu. 2022. AMOA: global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*, pages 7136–7146.

Ronghao Lin and Haifeng Hu. 2022. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 511–523.

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. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 2247–2256.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2023. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.*, 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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations*, pages 204–213.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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*, pages 2359–2369.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

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 *MM '22: The 30th ACM International Conference on Multimedia*, pages 3722–3729.

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*, page 6558.

Yao-Huzadeh2018multing Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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*.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew Arnold. 2021. Knowledge enhanced pretrained language models: A comprehensive survey. *arXiv preprint arXiv:2110.08455*.

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

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*, 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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13610–13624.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5634–5641.

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. 2018b. 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.

Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1262–1274.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767.

## A  Data Statistics and Analysis

To explore the sample distribution that each modality as dominant and verify our idea, we investigated four MSA benchmark datasets (MOSI, MOSEI, CH-SIMS, and CH-SIMSv2). The statistical and analytical results are as follows.

For MOSI and MOSEI, some researchers (Yu et al., 2020; Liu et al., 2022) have shown that the text modality is of higher importance and dominant. At the same time, some ablation studies (Hazarika et al., 2020; Lin and Hu, 2022; Yu et al., 2023) report about a 30% binary accuracy drop in these two datasets when removing the text modality (80%+ with text, while about 50%+ without text). Thus, the majority of samples in MOSI and MOSEI use text as the dominant modality.
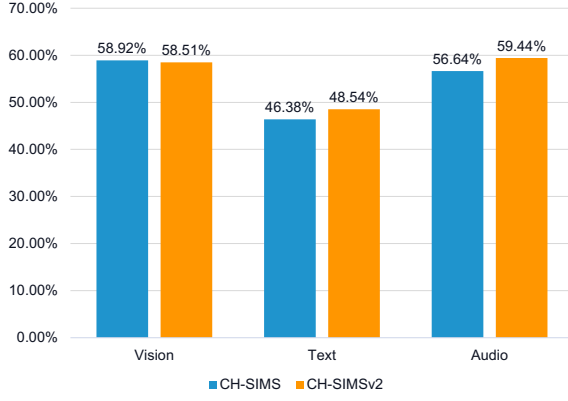
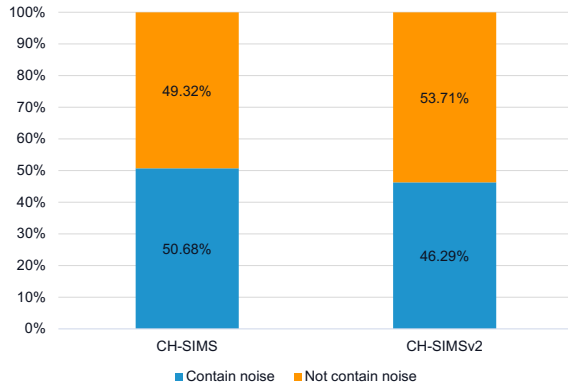Figure 7: The distribution where any modality as dominant.



Figure 8: The distribution where the sample contains the noise modality.

For CH-SIMS and CH-SIMSv2, we used the unimodal labels provided by the dataset to statistic the number of samples, as shown in Figure 7. We can observe that the number of samples dominated by vision, text, and audio modalities accounts for 45%-60% of the total, and the distribution of each modality is even. This indicates that each modality will be dominant, and this situation is not uncommon. It also shows that it is necessary to adjust the dominant modality dynamically. Notably, the sum of the proportions of three modalities in the same dataset is more than 100% because there may be multiple modalities that dominate in the same sample. Then, we statistic the number of samples containing noise modality (the sentiment polarity of this unimodal is different from that of the multimodal), as shown in Figure 8. We can see that the proportion of samples containing noise modality is around 50%, which further shows that the inability to dynamically adjust the contribution of each modality will limits the performance of MSA.

---

**Algorithm 1:** Training Process of KuDA

**Stage 1: Knowledge Inject Pretraining**
**Input:** External dataset $\varphi$ with the unimodal features $I_m$ and labels $y_m, m \in \{t, v, a\}$
**Output:** Pretrained adapters and decoders $\{\theta_m^{adapter}, \theta_m^{decoder} | m \in \{t, v, a\}\}$
**for** *each training epoch* **do**
    **for** *batch* $\big\{ \big(I_t^i, I_v^i, I_a^i\big)\big\}_{i=1}^N$ *from* $\varphi$ **do**
        Encode $I_m^i$ to $U_m^i$ as Eq. (1)-(4)
        Predict $\hat{y}_m^i$ using decoders as Eq. (5)
        Compute $\mathcal{L}_{reg}$ of each modality with $\hat{y}_m^i$ and $y_m^i$ as Eq. (14)
        Update parameters of encoding with knowledge injection module $\{\theta_m^{know} | m \in \{t, v, a\}\}$
    **end**
    Save $\{\theta_m^{adapter}, \theta_m^{decoder} | m \in \{t, v, a\}\}$ when achieves the best validate results
**end**
**Stage 2: Downstream Training**
**Input:** Target dataset $\mathcal{D}$ with the features $I_m, m \in \{t, v, a\}$ and labels $y$; Pretrained adapters and decoders
**Output:** Predictions $\hat{y}$
**for** *each training epoch* **do**
    **for** *batch* $\big\{ \big(I_t^i, I_v^i, I_a^i\big)\big\}_{i=1}^N$ *from* $\mathcal{D}$ **do**
        Encode $I_m^i$ to $U_m^i$ as Eq. (1)-(4)
        Predict $\hat{y}_m^i$ using decoders and calculate $R_m^i$ as Eq. (5), (6)
        Process dynamic attention fusion as Eq. (7)-(11)
        Predict $\hat{y}^i$ using MLP as Eq. (13)
        Compute $\mathcal{L}_{cor}, \mathcal{L}_{reg}, \mathcal{L}_{task}$ as Eq. (12), (14), (15)
        Update the model parameters except $\{\theta_m^{adapter}, \theta_m^{decoder} | m \in \{t, v, a\}\}$
    **end**
**end**

## B Training Process

KuDA uses a two-stage training method, which details are shown in Algorithm 1. In Stage 1, we pretrained the Encoding with Knowledge Injection module using external data. For CH-SIMS and CH-SIMSv2, we use the unimodal labels of the dataset itself for pretraining. For MOSI and MOSEI, considering the data scale, we translate the texts of CH-SIMS and CH-SIMSv2 into English and inject the sentiment knowledge of CH-SIMS into MOSI and that of CH-SIMSv2 into MOSEI. Notably, to compare fairly with the baselines, we only pretrain the task of unimodal sentiment prediction on all datasets and do not involve the MSA task. In Stage 2, to prevent the pretrained knowledge from being overwritten, we froze the Adapter and Decoder of each modality and performed the MSA task based on the pretrained knowledge.