MGCL: Modality-Granularity Collaborative Contrastive Learning for Multimodal Sentiment Analysis

Kai Liu ChuanQi Tao
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
Nanjing, China
{SX2416122, taochuanqi}@nuaa.edu.cn

Abstract

Multimodal sentiment analysis aims to more comprehensively recognize and understand human affective states by integrating text, audio, and visual information. Sentiment information includes both macro-level global sentiment categories, such as overall positive, and micro-level fine-grained sentiment signals, such as local cues conveyed through intonation, or frame-level expressions. Existing studies often treat text as dominant while marginalizing audio and visual modalities, underutilizing their rich auxiliary information. This causes models to over-rely on text, reducing their ability to fully understand sentiment and limiting robustness and generalization in complex scenarios. Moreover, current methods mainly focus on macro-level interactions, while overlooking the fine-grained information at the micro level. To this end, we propose a Modality-Granularity Collaborative Contrastive Learning framework (MGCL) to fully explore the commonalities and differences of sentiment across modalities and granularities. MGCL first extracts macro and micro features from each modality, and uses sentiment-intensity-based collaborative contrastive learning to capture similarities and differences between modalities and granular levels. In addition, to better utilize the complementary information across modalities and granularities, we introduce a cross modality-granularity fusion mechanism to enhance feature representations. Experimental results show that MGCL outperforms existing mainstream approaches on CH-SIMS, CMU-MOSI, and CMU-MOSEI.

Keywords: ***Contrastive learning, Cross modality-granularity fusion, Multimodal sentiment analysis, Sentiment intensity***

1. Introduction

In recent years, with the rise of short video–based social media platforms such as TikTok, Instagram Reels, and Snapchat Spotlight, and the rapid progress in deep learning techniques [1, 2, 3], sentiment analysis has evolved from traditional unimodal approaches to multimodal sentiment analysis (MSA). MSA leverages multiple modalities—such as text, audio, and vision—to compensate for the limitations of single-modal semantic expression, showing significant advantages in capturing complex emotional expressions, improving prediction accuracy, and modeling hierarchical structures of sentiment [4, 5, 6]. However, due to the inherent heterogeneity and hierarchical differences in sentiment expression across modalities, effectively modeling and fusing multi-granular sentiment information remains one of the challenges of current research.

Specifically, multimodal sentiment expression typically exhibits two core characteristics: modality heterogeneity and hierarchical granularity. Different modalities have distinct strengths in conveying sentiment: text is more suitable for expressing semantic-level emotions, audio reflects prosodic variations, and visual modalities are sensitive to micro-expressions and facial muscle movements. Sentiment information itself also possesses a multi-granular structure, with macro-level features corresponding to overall semantics or global emotional tendencies (*e.g.*, the positive or negative sentiment of a complete sentence), and micro-level features representing local details (*e.g.*, word-level intonation, prosodic fluctuations, or subtle facial expressions). In practice, the same sentiment may exhibit both redundancy and divergence across modalities and granularities—for instance, a mildly positive sentiment may be conveyed through a soft tone, while facial expression changes remain subtle. Such complementarity and divergence across modalities and granularities form a complex multi-dimensional interaction structure, posing higher requirements for multimodal sentiment analysis.

Most recent studies [7, 8] tend to prioritize text information as a more reliable source of sentiment. They usually project audio and visual modalities into a representation space close to the text modality and learn shared representations. However, this approach fails to fully exploit the auxiliary information provided by the audio and visual modali-

ties. Moreover, previous methods [9, 10] focused on exploring intra- and inter-modal interactions at the macro level, but failed to fully utilize the micro-granularity information of each modality. Furthermore, most approaches [8, 11] focus on a single aspect—either modality alignment or granularity fusion—while ignoring the deeper collaborative modeling between modality and granularity, which limits their ability to capture sentiment effectively. To address these limitations, we propose a Modality-Granularity Collaborative Contrastive Learning framework that jointly captures similarities, differences, and interactions across modalities and granularities, enabling more fine-grained sentiment modeling.

Specifically, we first pre-train macro- and micro-level feature encoders for each modality to extract emotional representations at different granularities. Then, we design a cross modality-granularity fusion mechanism, where macro-level features are used as queries and micro-level features as keys and values to enhance cross-granularity interactions. On this basis, we propose an sentiment-intensity-based modality-granularity collaborative contrastive learning strategy. It incorporates three contrastive learning schemes: same-modal and same-granularity, cross-modal and same-granularity, and same-modal and cross-granularity. These schemes guide the model to better distinguish subtle emotional differences. In addition, we introduce a single-modality prediction task as an auxiliary objective, and train the model using a multi-task objective function to improve its generalization and robustness. Our main contributions are as follows:

- We propose a modality-granularity dual-dimensional collaborative modeling approach that simultaneously captures modality heterogeneity and granularity hierarchy. Specifically, a dual-level feature extractor is employed to obtain both macro- and micro-level features for each modality. A cross-granularity attention mechanism is further introduced to enable information interaction between different granularities, effectively leveraging cross-modal complementarity and cross-granularity diversity.

- We propose a sentiment-intensity-based modality-granularity collaborative contrastive learning mechanism that integrates three contrastive strategies: same-modal and same-granularity, cross-modal and same-granularity, and same-modal and cross-granularity. By constructing semantically similar sample pairs that differ in either modality or granularity, the model is guided to learn fine-grained alignment across modalities and abstraction levels. In addition, we introduce a sample selection mechanism based on emotional intensity differences, which dynamically determines positive and negative pairs and assigns them differentiated weights. This enables the model to more sensitively capture subtle variations in emotional intensity during the contrastive learning process, thereby enhancing both the accuracy and discriminative capability of sentiment modeling.

- We conduct extensive experiments on three benchmark datasets: CH-SIMS, CMU-MOSI, and CMU-MOSEI. The experimental results demonstrate that our method consistently outperforms state-of-the-art approaches on the vast majority of evaluation metrics, validating the effectiveness and generalizability of our proposed model.

2. Related Work

2.1. Multimodal Sentiment Analysis

Multimodal sentiment analysis (MSA) aims to more accurately identify human emotional states by integrating information from multiple modalities such as text, audio, and vision. One of the core challenges in MSA is multimodal fusion, which seeks to integrate modality-specific representations into a unified and comprehensive representation. Zadeh *et al*. [12] proposed a novel neural architecture, the Memory Fusion Network (MFN), which effectively fuses multi-view information by combining view-specific LSTM systems with a multi-view gated memory. In an earlier work, Zadeh *et al*. [13] explicitly modeled unimodal, bimodal, and trimodal interaction dynamics and utilized a tensor fusion layer to capture complex inter-modal relationships. Tsai *et al*. [14] introduced a cross-modal attention mechanism along with a temporal alignment module to enhance multimodal information fusion. Hazarika *et al*. [4] proposed projecting each modality into modality-invariant and modality-specific subspaces, thereby effectively disentangling shared and unique features to enhance the fused representation. Rahman *et al*. [15] designed the Multimodal Adaptation Gate (MAG) to enable effective fine-tuning of BERT for multimodal tasks. Zhang *et al*. [11] introduced the Adaptive Hypermotion Learning Module (ALMT), which regulates visual and audio information using multi-scale linguistic features to suppress sentiment-irrelevant or conflicting information. Sun and Tian [10] designed a sequential cross-modal encoder to progressively fuse features proximate to and distant from the textual modality, thereby facilitating more effective feature integration.

However, these methods predominantly focus on either modality alignment or granularity modeling in isolation, overlooking the deeper synergy between modality and granularity dimensions.

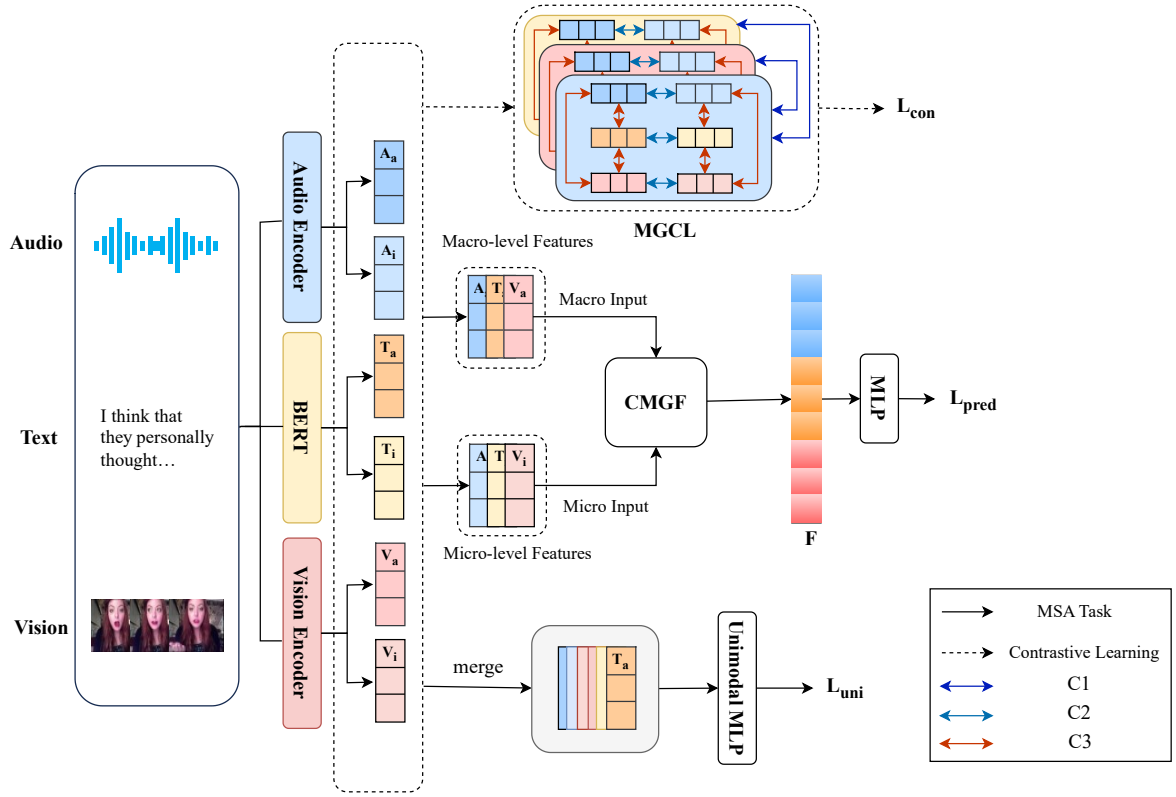

Figure 1. The overall structure of MGCL. (T_a, V_a, A_a) denote the macro-level features of the text, visual, and acoustic modalities, respectively, while (T_i, V_i, A_i) represent the corresponding micro-level features. C1, C2, C3 correspond to three contrastive learning strategies.

2.2. Contrastive Learning

Contrastive learning, as an unsupervised or weakly supervised representation learning paradigm, has been widely applied to multimodal tasks in recent years, achieving remarkable performance in learning effective representations [16, 17]. Existing contrastive learning approaches can generally be categorized into two types: self-supervised contrastive learning [16, 18, 19] and supervised contrastive learning [20, 21], with the primary distinction being whether or not label information is used to construct positive and negative pairs.

In recent years, supervised contrastive learning has attracted increasing attention in multimodal sentiment analysis (MSA). Yu *et al*. [22] proposed a self-supervised multimodal sentiment analysis framework that incorporates consistency loss and contrastive loss to enhance the collaborative representation across modalities while preserving modality-specific information. Mai *et al*. [9] introduced Hycon, the first framework to apply contrastive learning to modality interaction modeling. By constructing positive and negative pairs across modalities, Hycon effectively learns joint representations from text, image, and audio data. Yang *et al*. [7] proposed ConFEDE, a unified framework for contrastive representation learning and contrastive feature decomposition. In this framework, each modality is decomposed into similarity and dissimilarity features, and contrastive relations are constructed using the similarity features of text as anchors, thereby guiding the model to capture both the consistency and divergence across modalities. Yu *et al*. [23] proposed ConKI, a contrastive knowledge injection model that learns both domain-specific and general knowledge representations for MSA. This method utilizes a hierarchical contrastive learning process to enhance emotional representation and improve sentiment prediction performance. Sun and Tian [10] introduced SFTTR, which decomposes each modality into Text-close and Text-far representations and uses contrastive learning to explore semantic similarities and differences between text and other modalities.

Despite the promising progress of these approaches in modality-level contrastive learning, they primarily focus on aligning and distinguishing modality-specific representations, while failing to thoroughly model the shared and distinctive characteristics of emotions across different granularities. Moreover, prior semantic information, such as emotional intensity, has not been fully exploited.

To address these limitations, this work proposes a collaborative modality-granularity contrastive learning mech-

anism, which jointly models fine-grained relationships across modalities and granularities, enabling more precise and nuanced sentiment representation learning.

3. Method

3.1. Overall Architecture

The overall architecture of MGCL is illustrated in Figure 1. Given an input sample, we first perform representation modeling for each modality using corresponding modality-specific encoders. Unlike previous approaches, we design and pretrain separate models to extract both macro- and micro-level features for each modality. Specifically, for the textual modality, we adopt BERT [24] as the text encoder. The [CLS] token representation is used to capture the macro-level feature of the text (i.e., T_{macro}), while the token-level hidden states from the final layer serve as the micro-level features (i.e., T_{micro}). To enhance the micro-level representation, we further apply operations such as temporal average pooling, adjacent token differencing, and feature concatenation on the token embeddings.

For the visual modality, we utilize a Transformer encoder [25] to model the sequence of video frames. An attention pooling mechanism is applied to obtain a global semantic representation as the macro-level feature (i.e., V_{macro}), while the hidden states corresponding to individual frames are retained as the micro-level features (i.e., V_{micro}). Similar to the textual modality, we enhance the visual micro-level representations through average pooling, inter-frame differencing, and feature concatenation. For the acoustic modality, we adopt the same architecture and processing strategy as used for the visual modality, resulting in macro-level (i.e., A_{macro}) and micro-level (i.e., A_{micro}) representations of the audio signals.

Subsequently, the macro- and micro-level features of each modality are fed into separate projection modules to map them into a unified representation space. Specifically, the macro-level projection module consists of layer normalization, a linear layer with Tanh activation, and a Dropout layer. In contrast, the micro-level projection module includes layer normalization, a hidden linear layer with ReLU activation, a subsequent linear layer with Tanh activation, and a Dropout operation. Through these projection operations, we obtain the unified macro-level features (i.e., T_a, V_a, A_a) and micro-level features (i.e., T_i, V_i, A_i) in a shared representation space.

After obtaining the six projected features, we feed them into two key modules: the Modality-Granularity Collaborative Contrastive Learning (MGCL) and the Cross Modality-Granularity Fusion module (CMGF), which are designed to explore fine-grained alignment and complementary relationships among multimodal features.

Following the design of Yang *et al*. [7], we apply unimodal sentiment prediction as an auxiliary task, where the macro- and micro-level features of each modality are individually used for sentiment classification. This auxiliary task aims to enhance intra-modal semantic modeling and improve the model's robustness in scenarios involving modality degradation or missing modalities.

These modules will be described in detail in the following sections. Finally, we optimize the entire model using a multi-task joint learning strategy, which simultaneously minimizes the contrastive learning loss, the fusion-based prediction loss, and the unimodal prediction losses, thereby enabling effective learning from multiple granularities and perspectives.

3.2. Modality-Granularity Collaborative Contrastive Learning

When selecting positive and negative sample pairs, we observe that most existing contrastive learning methods assume equal importance for all sample pairs throughout training, without adequately considering how differences in sentiment intensity affect the structure of the representation space. We argue that in sentiment modeling tasks, samples with varying sentiment intensity (i.e., label value differences) should exhibit different levels of proximity in the representation space.

For example, a sample pair (i, j) with a label difference of 0.1 should be represented closer together than a pair (i, k) with a difference of 0.5, and thus is more appropriately treated as a positive pair. Conversely, for negative pairs, greater sentiment intensity differences should correspond to a larger separation in the representation space. Following the sentiment intensity-guided strategy proposed by Yang *et al*. [8], we design a weighting mechanism for sample pair construction, where the contribution of each pair is modulated according to the emotional discrepancy between them. During contrastive loss computation, we assign higher weights to sample pairs with larger intensity differences. This approach reinforces the model's sensitivity to sentiment boundaries and enhances its discriminative ability in distinguishing subtle sentiment variations.

First, we introduce a sentiment intensity difference threshold δ, which is used to determine the relationship between sample pairs. Specifically, if the absolute difference between the sentiment labels of two samples exceeds δ, the pair is treated as a negative pair. Otherwise, it is considered a positive pair. In our experiments, we set $\delta = 0.4$, a value empirically determined through preliminary experiments to achieve optimal performance.

$$\begin{cases} E_{(i,j)} \leq \delta, & (i,j) \in \text{positive pairs} \\ E_{(i,j)} > \delta, & (i,j) \in \text{negative pairs} \end{cases} \tag{1}$$

Where $E_{(i,j)}$ represents the difference in sentiment intensity between sample pairs (i, j).

Subsequently, to more effectively distinguish sample pairs with large sentiment intensity differences in the representation space, we introduce a sentiment intensity difference-based weighting mechanism, which adaptively adjusts the contribution of each sample pair to the contrastive loss. As shown in Equation (2), both positive and negative pairs are assigned contrastive weights according to the magnitude of their label differences. This guides the model to draw emotionally similar sample pairs closer, while pushing apart those with significant sentiment discrepancies, thereby enhancing the emotional discriminability of the learned representations.

$$\begin{cases} 1.5 \times \left| \tanh(E_{(i,j)}) \right|, & (i,j) \in \text{negative pairs} \\ 1.5 \times \left| \tanh(E_{(i,j)} - \delta) \right|, & (i,j) \in \text{positive pairs} \end{cases} \tag{2}$$

Considering that sentiment intensity labels may vary in scale across different datasets, we normalize all label values to the range$[-1, 1]$ to ensure numerical stability during contrastive learning and enhance the generalization ability across datasets. Specifically, given a batch B containing multiple sample pairs, we compute the normalized sentiment intensity difference for each pair (i, j) as follows:

$$E_{(i,j)} = \left| y_i - y_j \right|, \quad i \neq j, \quad j \in B \tag{3}$$

Where y_i and y_j denote the sentiment intensity labels of samples i and j, respectively.

Based on the aforementioned sentiment intensity difference criterion, we first construct initial sets of positive and negative sample pairs for each sample i. Building upon this, we introduce a carefully designed modality-granularity collaborative contrastive strategy to more comprehensively explore the commonalities and differences of multimodal emotional features across both modalities and granularities.

In the same-modal and same-granularity contrastive strategy, our goal is to enhance the robustness and discriminability of feature representations within each modality. Specifically, we select macro-level features as the contrastive objects and construct positive pairs from different samples within the same modality that share identical sentiment labels. For example, in the textual modality, macro features T_a^i and T_a^j from two samples with the same sentiment label are contrasted to pull their representations closer in the embedding space. This strategy helps the model better perceive similar emotional expressions and improves semantic clustering. The construction of positive and negative sample pairs is defined as follows:

$$P^i_{\text{same,same}} = \left\{ \left(T_a^i, T_a^j\right), \left(V_a^i, V_a^j\right), \left(A_a^i, A_a^j\right) \;\middle|\; (i,j) \in \text{positive pairs} \right\} \tag{4}$$

$$N^i_{\text{same,same}} = \left\{ \left(T_a^i, T_a^k\right), \left(V_a^i, V_a^k\right), \left(A_a^i, A_a^k\right) \;\middle|\; (i,k) \in \text{negative pairs} \right\} \tag{5}$$

In the cross-modal and same-granularity contrastive strategy, we focus on ensuring consistency of representations across different modalities for the same semantic content. This strategy also centers on macro-level features. Specifically, for the same sample, features from different modalities (*e.g.*, T_a^i and V_a^i) are treated as positive pairs for contrastive learning, guiding the model to align semantically similar cross-modal information in the latent representation space. This approach helps mitigate distribution discrepancies caused by modality heterogeneity and enhances the consistency and fusion of cross-modal emotional features. The construction of positive and negative sample pairs is defined as follows:

$$\begin{aligned} P^i_{\text{cross,same}} = \big\{ & (T_a^i, A_a^i), (T_a^i, V_a^i), (V_a^i, A_a^i) \big\} \\ & \cup \big\{ (T_a^i, A_a^j), (T_a^i, V_a^j), (V_a^i, A_a^j), \\ & (A_a^i, T_a^j), (V_a^i, T_a^j), (A_a^i, V_a^j) \\ & \mid (i,j) \in \text{positive pairs} \big\} \end{aligned} \tag{6}$$

$$\begin{aligned} N^i_{\text{cross,same}} = \big\{ & (T_a^i, V_a^k), (V_a^i, A_a^k), (T_a^i, A_a^k), \\ & (V_a^i, T_a^k), (A_a^i, V_a^k), (A_a^i, T_a^k) \\ & \mid (i,k) \in \text{negative pairs} \big\} \end{aligned} \tag{7}$$

In the same-modal and cross-granularity contrastive strategy, our aim is to establish representational connections between macro- and micro-level features, promoting consistency and complementarity across semantic hierarchies. Taking the textual modality as an example, we pair the macro-level feature T_a^j with the micro-level feature T_i^j from the same sample, guiding the model to build collaborative relationships across different granularities. This strategy enhances the model's sensitivity to fine-grained emotional information while reinforcing structural alignment and semantic complementarity between features at different levels. The construction of positive and negative sample pairs is defined as follows:

$$P^i_{\text{same,cross}} = \left\{ \left(T_a^i, T_i^i\right), \left(V_a^i, V_i^i\right), \left(A_a^i, A_i^i\right) \;\middle|\; (i,i) \in \text{positive pairs} \right\} \tag{8}$$

$$N^i_{\text{same,cross}} = \left\{ \left(T_a^i, T_i^k\right), \left(V_a^i, V_i^k\right), \left(A_a^i, A_i^k\right) \;\middle|\; (i,k) \in \text{negative pairs} \right\} \tag{9}$$

It is worth noting that we intentionally do not introduce a direct cross-modal and cross-granularity contrastive strategy. Although such a setting may appear intuitive for modeling modality–granularity interactions, it involves simultaneous shifts in both modality and semantic granularity, which can lead to semantic misalignment and ambiguous supervision signals in contrastive learning. In particular, representations from different modalities and different granularity levels often exhibit heterogeneous abstraction structures and modality-specific biases, making it difficult to define semantically comparable positive and negative pairs.

Directly enforcing contrastive constraints across both dimensions may introduce noisy alignments and adversely affect optimization stability.

By combining the same-modal and same-granularity, cross-modal and same-granularity, and same-modal and cross-granularity pairs of sample i, we obtain the sets of positive and negative sample pairs P_i and N_i for sample i during the contrastive learning process, as follows:

$$P_i = P^i_{\text{same,same}} \cup P^i_{\text{cross,same}} \cup P^i_{\text{same,cross}} \tag{10}$$

$$N_i = N^i_{\text{same,same}} \cup N^i_{\text{cross,same}} \cup N^i_{\text{same,cross}} \tag{11}$$

Given a batch B, our collaborative contrastive loss L_{con} is computed as follows:

$$L_{con} = -\sum_{i \in B} \log \frac{\sum_{(a,p) \in P^i} \exp\left(\frac{W_{(i,j)} \cdot \text{sim}(\alpha,p)}{\tau}\right)}{\sum_{(a,q) \in P^i \cup N^i} \exp\left(\frac{W_{(i,j)} \cdot \text{sim}(\alpha,q)}{\tau}\right)} \tag{12}$$

where $W_{(i,j)}$ is the corresponding weight for the sample pair (i, j).

3.3. Cross Modality-Granularity Fusion

From the projection operations described above, we obtain both macro- and micro-level features for each modality. However, simply concatenating these features may overlook deep semantic interactions across modalities and granularities, and may introduce information redundancy, which could negatively impact downstream task performance. Inspired by [10], we propose a Cross Modality-Granularity Fusion (CMGF) mechanism to fully exploit the complementary information between modalities and granularities, thereby enhancing the interaction and collaborative representation of multimodal features. The structure of this mechanism is illustrated in Figure 2. Specifically, we use the macro-level feature of each modality as a query, and the micro-level features of the other two modalities as keys and values, enabling dynamic cross-modal and cross-granularity interaction via a multi-head attention mechanism. For example, in the case of the textual modality, the macro feature T_a interacts with the micro-level features of the visual (V_i) and acoustic (A_i) modalities to capture fine-grained semantic cues from other modalities. Similarly, Va and Aa interact with the micro-level features of the remaining modalities to enrich their own representations with complementary information. This process is formalized as follows:

$$F^1_{en} = MTM(T_a, V_i, A_i) \tag{13}$$

$$F^2_{en} = MTM(V_a, T_i, A_i) \tag{14}$$

$$F^3_{en} = MTM(A_a, T_i, V_i) \tag{15}$$

Where $\text{MTM}(\cdot)$ represents the fusion module.

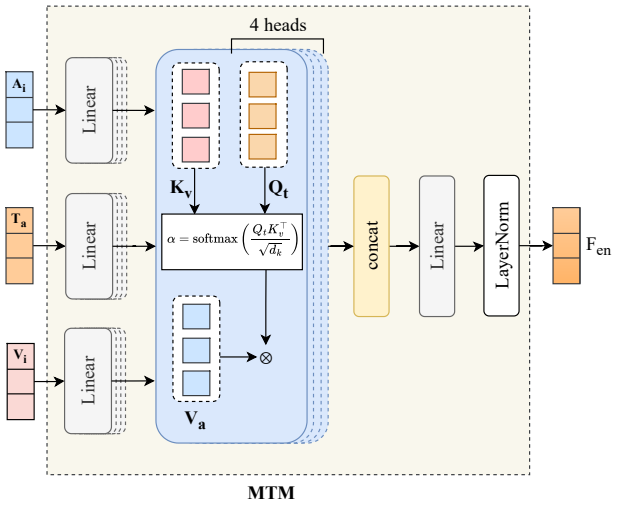

Figure 2. Structure of the CMGF module, illustrated with the textual modality as an example.

Finally, we concatenate the three enhanced macro-level features to construct a unified fusion representation, which is then fed into a three-layer multilayer perceptron (MLP) as the sentiment intensity predictor to perform sentiment intensity regression.

$$F = \begin{bmatrix} F^1_{en} \\ F^2_{en} \\ F^3_{en} \end{bmatrix} \tag{16}$$

3.4. Overall Learning Objectives

After multimodal fusion, we employ an MLP with ReLU activation functions as the classifier to obtain the final prediction results. This choice is primarily motivated by the MLP's ability to capture complex nonlinear relationships in the input data—an essential characteristic for sentiment analysis tasks, where latent patterns and subtle variations are often highly nonlinear and multifaceted. We use the joint multimodal representation F as the input to the classifier. Let the batch of samples be denoted as B. For a given sample $i \in B$, let its predicted sentiment intensity from the classifier be $\hat{y}^i$. The multimodal prediction loss is then computed using the mean squared error (MSE) as follows:

$$\hat{y}^i = \text{MLP}(F) \tag{17}$$

$$L_{\text{pred}} = \frac{1}{n} \sum_{i=1}^{n} (y^i - \hat{y}^i)^2 \tag{18}$$

Where n is the number of samples in the batch and y^i is the multimodal label.

In addition, for each sample i, we feed the obtained macro- and micro-level features $\left[T^i_a, V^i_a, A^i_a, T^i_i, V^i_i, A^i_i\right]$ into separate MLP classifiers as auxiliary tasks, resulting

Dataset	#Train	#Valid	#Test	#Total	Language
CH-SIMS	1,368	456	457	2,281	Chinese
CMU-MOSI	1,284	229	686	2,199	English
CMU-MOSEI	16,326	1,871	4,659	22,856	English

Table 1. The statistics of CH-SIMS, CMU-MOSI and CMU-MOSEI.

in six individual predictions denoted $\hat{u}^i$. Specifically, the unimodal prediction loss is computed as follows:

$$\hat{u}^i = \text{MLP}([T_a^i, V_a^i, A_a^i, T_i^i, V_i^i, A_i^i]) \tag{19}$$

$$u^i = [y^i, y^i, y^i, y_T^i, y_V^i, y_A^i] \tag{20}$$

$$L_{\text{uni}} = \frac{1}{n} \left\| u^i - \hat{u}^i \right\|_2^2 \tag{21}$$

where the vector $u^i = \left[y^i, y^i, y^i, y_T^i, y_V^i, y_A^i\right]$ represents the ground-truth labels for unimodal prediction. In other words, each decomposed feature is individually regularized to perform a separate prediction task.

The overall loss function can be formulated as follows:

$$L_{\text{all}} = L_{\text{pred}} + \alpha L_{\text{con}} + \beta L_{\text{uni}} \tag{22}$$

where L_{con} is the collaborative contrastive loss, L_{pred} is the multimodal prediction loss, and L_{uni} is the auxiliary unimodal prediction loss. The coefficients α and β are hyperparameters that balance the contributions of each loss term.

4. Experiment

4.1. Dataset and Metrics

We conducted extensive experiments on three publicly available benchmark datasets: CMU-MOSI [26], CMU-MOSEI [27], and CH-SIMS [28]. The statistic details of three datasets are shown in Table 1.

In our experiments, we report the averaged results over three runs with different random seeds, evaluating overall performance on both classification and regression tasks. For classification, we assess multi-class accuracy and weighted F1 scores. This includes binary (Acc-2), three-class (Acc-3), and five-class (Acc-5) accuracy on the CH-SIMS dataset, as well as binary and seven-class (Acc-7) accuracy on the CMU-MOSI and CMU-MOSEI datasets. For the binary classification results (Acc-2 and F1 scores) on CMU-MOSI and CMU-MOSEI, we evaluate using two different label partitioning schemes: negative/non-negative (retaining the neutral category) [13, 22] and negative/positive (excluding the neutral category) [14, 22]. Regarding regression, we report two metrics: mean absolute error (MAE) and Pearson correlation coefficient (Corr). Except for MAE, higher values of all other metrics indicate better model performance.

Stage	Dataset	Model / Encoder	Learning Rate	Epochs	Batch Size
Unimodal	CH-SIMS	bert-base-chinese	1e-5	150	64
	CMU-MOSI	bert-base-uncased	1e-5	150	64
	CMU-MOSEI	bert-base-uncased	1e-5	150	64
	CH-SIMS	1-layer Transformer (V/A)	1e-4	200	128
	CMU-MOSI	1-layer Transformer (V/A)	1e-4	200	128
	CMU-MOSEI	3-layer Transformer (V/A)	1e-4	200	128
Multimodal	CH-SIMS	MGCL	1e-4	50	64
	CMU-MOSI	MGCL	1e-4	50	64
	CMU-MOSEI	MGCL	1e-4	25	8

Table 2. Experimental Settings and Hyperparameters.

4.2. Experimental Settings

To ensure a fair comparison with existing benchmarks, we adopt standard experimental protocols widely used in recent competitive and state-of-the-art studies. The training process consists of two stages: a unimodal training stage and a multimodal training stage.

In the unimodal training stage, we fine-tune bert-base-chinese for CH-SIMS and bert-base-uncased for CMU-MOSI and CMU-MOSEI to obtain textual representations. Subsequently, Transformer encoders are adopted as the Vision Encoder and Audio Encoder, as illustrated in Figure 1. Specifically, for CH-SIMS and CMU-MOSI, single-layer Transformer encoders are employed to model the visual and acoustic modalities, respectively. For CMU-MOSEI, deeper Transformer encoders with three layers are used to accommodate its larger scale. In the multimodal training stage, MGCL is trained for multimodal sentiment analysis by leveraging the pretrained unimodal encoders. All experiments are conducted on a single NVIDIA RTX 4090 GPU, and the detailed experimental settings and hyperparameters are reported in Table 2.

4.3. Baseline

To comprehensively evaluate the performance of our MGCL model, we conducted a fair comparison with several advanced and state-of-the-art baseline methods. These baselines include LF-DNN [28], MFN [12], LMF [29], TFN [13], MulT [14], MISA [4], MAG-BERT [15], Self-MM [22], ConFEDE [7] and SFTTR [10].

4.4. Results

Tables 3 and 4 summarize the performance comparisons of all methods on the CH-SIMS, CMU-MOSI, and CMU-MOSEI datasets. We report the average results of our proposed model over three independent runs with different random seeds to reduce randomness and improve result stability. For each evaluation metric, the best results are highlighted in bold font for clarity.

On the CH-SIMS dataset, our proposed method outperforms nearly all mainstream baseline models across almost all evaluation metrics. Compared to the best-performing contrastive method, SFTTR, we achieve significant improvements on several key metrics: a 1.11% increase in

Model	CH-SIMS					
	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
LF-DNN	78.87	79.87	66.91	41.62	0.420	0.612
MFN	77.90	77.88	65.73	39.47	0.435	0.582
LMF	77.77	77.88	64.68	40.53	0.441	0.576
TFN	78.38	78.62	65.12	39.30	0.432	0.591
MulT	78.56	79.66	64.77	37.94	0.453	0.561
MISA	76.54	76.59	-	-	0.447	0.563
Self-MM*	78.71	78.76	65.47	42.94	0.411	0.601
ConFEDE*	79.26	79.47	68.82	**45.60**	0.389	0.646
SFTTR*	80.58	80.49	69.15	45.51	**0.384**	0.649
MGCL	**81.69**	**81.66**	**69.88**	43.40	0.402	**0.664**

Table 3. Results on CH-SIMS. Models with * were replicated under the same conditions, while other results are from published papers or official repositories.

Acc-2, a 1.17% improvement in F1 score, a 0.73% gain in Acc-3, and a 1.5% rise in Pearson correlation (Corr). These results demonstrate that our approach possesses stronger modeling capabilities in capturing the consistency of emotional expressions and the synergy among modalities.

However, our model performs slightly worse than some existing methods on the Acc-5 and MAE metrics. These metrics require finer-grained sentiment discrimination and more accurate modeling of continuous sentiment intensity, and are therefore more sensitive to subtle representation biases than coarse-grained evaluations. We believe this phenomenon may be attributed to the following two factors. (1) Characteristics inherent to the fusion mechanism. When incorporating micro-level features from other modalities, certain samples may introduce noise or misleading cues, particularly when emotional expressions across modalities are inconsistent. While such noise may have limited influence on coarse-grained classification performance, it can blur fine-grained decision boundaries and distort continuous intensity estimation, leading to more noticeable degradation on Acc-5 and MAE. In addition, the fusion module may exhibit a smoothing effect on representations of samples with extreme emotions, which reduces sensitivity to large sentiment intensity variations and hinders discrimination among adjacent fine-grained categories. (2) Limitations of the sentiment-intensity-guided contrastive strategy. Although contrastive learning is guided by sentiment intensity differences, the current linear distance-based formulation may not fully capture the nonlinear psychological perception underlying human emotion annotation. As a result, differences that are perceptually uneven may be treated uniformly, which can negatively affect continuous prediction and fine-grained evaluation metrics such as MAE. Overall, these observations suggest that samples with strong cross-modal inconsistency or extreme sentiment intensity pose greater challenges for fine-grained sentiment modeling, helping to explain why performance differences are more apparent on Acc-5 and MAE than on coarse-grained metrics.

On the CMU-MOSI dataset, our proposed MGCL method outperforms all existing baseline models on several key metrics. Specifically, MGCL achieves the best performance in terms of Pearson correlation (Corr), Acc-2, and F1 score under both the negative/non-negative (NN) and negative/positive (NP) partition settings. Additionally, our method surpasses most competing models in Acc-7 and MAE metrics, demonstrating strong overall stability.

These results indicate that the modality-granularity collaborative contrastive learning mechanism proposed in MGCL effectively models the interactions and expressions of multimodal features, maintaining excellent robustness and generalization even on relatively small-scale datasets like CMU-MOSI. On the CMU-MOSEI dataset, MGCL attains leading results across all metrics except for Acc-7 and MAE, further validating the adaptability and effectiveness of our approach on large-scale, diverse sentiment data.

We observe that MGCL achieves significantly superior Pearson correlation (Corr) scores across the CH-SIMS, CMU-MOSI, and CMU-MOSEI datasets. This trend indicates that MGCL has a strong capability to model the relative variations in sentiment intensity. We attribute this advantage primarily to the modality-granularity collaborative contrastive learning mechanism introduced in MGCL: by constructing multidimensional sample pairs and performing contrastive optimization guided by sentiment intensity differences, the model effectively learns the relative emotional distances between samples, thereby enhancing its sensitivity to sentiment intensity ranking. Unlike traditional regression models that focus solely on predicting absolute labels, MGCL explicitly optimizes the representation distribution of sample pairs during training. This encourages semantically similar samples to be closer in the embedding space while samples with large sentiment intensity differences are pushed farther apart. Such an optimization naturally fosters a more coherent emotional ranking structure, ultimately reflected in higher Corr scores during evaluation.

4.5. Ablation Study

To evaluate the impact of each key component in our proposed model on overall performance, we conducted systematic ablation studies on the CH-SIMS dataset. The results are presented in Table 5. These ablation settings are designed to analyze how each component contributes to specific learning objectives, such as cross-modal alignment, cross-granularity consistency, and modality-specific sentiment modeling.

Specifically, we individually removed the following components or strategies: “w/o CL”: Removing the contrastive learning module to verify its contribution to feature representation and sentiment recognition; “w/o C3”: Re-

Model	CMU-MOSI					CMU-MOSEI				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr
LF-DNN	77.52/78.63	77.46/78.63	34.52	0.955	0.658	80.60/82.74	80.85/82.52	50.83	0.580	0.709
MFN	77.4/-	77.3/-	34.1	0.965	0.632	78.94/82.86	79.55/82.85	51.34	0.573	0.718
LMF	-/82.5	-/82.4	33.2	0.917	0.695	80.54/83.48	80.94/83.36	51.59	0.576	0.717
TFN	-/80.8	-/80.7	34.9	0.901	0.698	78.50/81.89	78.96/81.74	51.60	0.573	0.714
MulT	-/83.0	-/82.8	40.0	0.871	0.698	81.15/84.63	81.56/84.52	52.84	0.559	0.733
MISA	81.8/83.4	81.7/83.6	42.3	0.783	0.776	82.59/84.23	82.67/83.97	52.20	0.568	0.724
MAG-BERT	82.13/83.54	81.12/83.58	41.43	0.790	0.766	82.51/84.82	82.77/84.71	50.41	0.583	0.741
Self-MM*	82.51/84.7	82.6/84.91	45.79	0.712	0.792	82.68/84.96	82.95/84.93	53.46	**0.529**	0.767
ConFEDE*	82.62/84.45	82.6/84.48	44.92	0.731	0.792	81.21/85.75	81.77/85.74	53.14	0.540	0.766
SFTTR*	81.92/83.85	81.92/83.9	**45.79**	**0.712**	0.793	82.47/85.82	82.84/85.63	**53.69**	0.531	0.761
MGCL	**83.04/84.96**	**82.96/84.94**	45.19	0.727	**0.794**	**83.12/85.85**	**83.33/85.66**	52.79	0.542	**0.770**

Table 4. Results on CMU-MOSI and CMU-MOSEI. All baseline settings and results are consistent with those in Table 3. For Acc-2 and F1 scores, the values to the left of the "/" correspond to the negative/non-negative (NN) setting, while those to the right correspond to the negative/positive (NP) setting.

moving the same-modal and cross-granularity contrastive strategy while retaining the other contrastive terms, to assess the effectiveness of granularity-collaborative contrastive learning; "w/o uni": Removing the unimodal prediction module to investigate the influence of auxiliary supervision in multi-task learning; "w/o fusion": Replacing our designed cross-modal granularity fusion module with simple feature concatenation to validate the effectiveness of the proposed interaction and fusion structure.

The experimental results indicate that removing any of the key components leads to a decline in model performance, confirming the importance of each module within the overall framework. Among the four ablation settings, removing the fusion module ("w/o fusion") has the greatest impact, causing a 1.33% drop in Acc-2 compared to the full model. This strongly demonstrates the effectiveness of our proposed cross modality-granularity fusion mechanism in enhancing feature representation and improving sentiment recognition capability. Specifically, without the proposed fusion module, the model is unable to explicitly optimize cross-modal semantic alignment and fine-grained interaction across different granularities, leading to weaker joint representations for sentiment prediction.

Model	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
w/o CL	79.94	79.98	69.15	42.16	0.409	0.650
w/o C3	81.40	81.31	69.29	43.25	0.407	0.659
w/o uni	81.40	81.17	69.15	43.39	0.414	0.646
w/o fusion	79.36	79.18	68.49	42.45	0.405	0.662
MGCL	**81.69**	**81.66**	**69.88**	**43.40**	**0.402**	**0.664**

Table 5. Ablation study results on CH-SIMS. Each score represents the average over three runs.

Furthermore, when the entire contrastive learning module ("w/o CL") is removed, the model's performance drops significantly. This degradation suggests that removing contrastive learning weakens the explicit optimization of cross-modal and cross-granularity alignment objectives, resulting in less discriminative sentiment representations. In comparison, removing the same-modal cross-granularity contrastive component ("w/o C3") also leads to some performance degradation, but the impact is relatively smaller. This suggests that granularity-collaborative contrastive learning serves as a complementary pathway that enhances the model's sensitivity to fine-grained emotional details. In particular, this component enforces consistency between macro- and micro-level representations within the same modality, which directly supports the optimization of fine-grained sentiment modeling.

When the unimodal prediction module is removed (w/o uni), the model performance drops accordingly. This result indicates that different modalities in the CH-SIMS dataset contain modality-specific sentiment cues. Incorporating unimodal prediction as an auxiliary task helps the model capture these unique representations more effectively, thereby improving overall performance. Removing this auxiliary task weakens the modality-specific supervision in the multi-task learning objective, which in turn reduces the quality of the fused representation.

To further evaluate the effectiveness of contrastive learning guided by sentiment intensity, we present ablation results in Table 6. Here, "w/o CL" indicates the model trained without contrastive learning, while "w/o weight" refers to using sentiment labels to select positive and negative pairs without incorporating sentiment-intensity weights.

The experimental results show that contrastive learning guided by sentiment intensity substantially enhances performance on the CH-SIMS dataset. In contrast, the "w/o

weight" setting leads to a noticeable performance drop. This degradation can be attributed to the fact that SIMS includes fine-grained sentiment intensity labels. Therefore, without applying sentiment-intensity weights, contrastive learning struggles to capture nuanced sentiment distinctions, thereby limiting overall model performance. This indicates that sentiment-intensity weighting is essential for effectively optimizing fine-grained contrastive objectives on datasets with continuous sentiment annotations. By integrating weights derived from sentiment intensity, our proposed contrastive learning method achieves significant improvements across all evaluation metrics. These findings demonstrate the effectiveness of sentiment-intensity-guided contrastive learning in enhancing model performance.

Overall, the ablation results demonstrate that each module in MGCL is closely associated with a specific optimization objective, and the observed performance degradation in each ablation setting is consistent with the corresponding objective being weakened.

Model	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
w/o CL	79.94	79.98	69.15	42.16	0.409	0.650
w/o weight	80.63	80.01	69.52	42.56	0.410	0.652
MGCL	**81.69**	**81.66**	**69.88**	**43.40**	**0.402**	**0.664**

Table 6. Ablation study of sentiment-intensity-guided contrastive learning on CH-SIMS.

5. Conclusion

In this paper, we propose MGCL, a modality-granularity contrastive learning framework for multimodal sentiment analysis. MGCL first employs pretrained encoders to extract macro-level features (global semantics) and micro-level features (fine-grained dynamics) for each modality separately. Subsequently, a cross modality-granularity fusion mechanism is employed, where micro-level features are leveraged to enhance macro-level representations, thereby strengthening the modeling of sentiment expressions. To effectively capture cross-modal consistency and cross-granularity complementarity, MGCL introduces a modality-granularity collaborative contrastive learning guided by sentiment intensity. Additionally, unimodal prediction is incorporated as an auxiliary task to improve the model's robustness and generalization capability. Experimental results on three mainstream multimodal sentiment analysis datasets—CH-SIMS, CMU-MOSI, and CMU-MOSEI—demonstrate that MGCL significantly outperforms existing state-of-the-art methods on multiple key metrics, particularly exhibiting superior ability in modeling sentiment intensity as reflected by the Pearson correlation coefficient (Corr).

Limitations

Although our proposed MGCL method achieves promising results in multimodal sentiment analysis, several limitations need to be considered. First, our approach is specifically designed for scenarios where all modalities—text, video, and audio—are available. The model's performance may degrade when one or more modalities are missing. Additionally, since we pretrain separate encoders to extract both macro and micro features for each modality, this approach demands more time and computational resources during training on large-scale datasets. Furthermore, the model deployment phase requires increased GPU memory consumption due to the dual-encoder architecture for each modality. In future work, we will further explore modal robust learning, adaptive fusion strategies, and knowledge distillation to effectively address the aforementioned limitations on the basis of the existing framework.

References

[1] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox. COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 22605–22618. Curran Associates, Inc., 2020. 1

[2] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu. Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7327–7337, Nashville, TN, USA, June 2021. IEEE. 1

[3] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online, Nov. 2020. Association for Computational Linguistics. 1

[4] D. Hazarika, R. Zimmermann, and S. Poria. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, Seattle WA USA, Oct. 2020. ACM. 1, 2, 7

[5] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651, Lisboa Portugal, Oct. 2022. ACM. 1

[6] X. Zhang, W. Wei, and S. Zou. Modal Feature Optimization Network with Prompt for Multimodal Sentiment Analysis. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4611–4621, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. 1

[7] J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment

Analysis. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630, Toronto, Canada, July 2023. Association for Computational Linguistics. 1, 3, 4, 7

[8] Y. Yang, X. Dong, and Y. Qiang. 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, Mexico City, Mexico, 2024. Association for Computational Linguistics. 1, 2, 4

[9] S. Mai, Y. Zeng, S. Zheng, and H. Hu. Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289, July 2023. 2, 3

[10] K. Sun and M. Tian. Sequential Fusion of Text-close and Text-far Representations for Multimodal Sentiment Analysis. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 40–49, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. 2, 3, 6, 7

[11] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, Singapore, Dec. 2023. Association for Computational Linguistics. 2

[12] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency. Memory Fusion Network for Multi-view Sequential Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. 2, 7

[13] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 2, 7

[14] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. 2, 7

[15] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque. Integrating Multimodal Information in Large Pretrained Transformers. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online, July 2020. Association for Computational Linguistics. 2, 7

[16] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *Advances in Neural Information Processing Systems*, volume 34, pages 24206–24221. Curran Associates, Inc., 2021. 3

[17] C. Liu, Y. Fu, C. Xu, S. Yang, J. Li, C. Wang, and L. Zhang. Learning a Few-shot Embedding Model with Contrastive Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8635–8643, May 2021. Number: 10. 3

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, Nov. 2020. ISSN: 2640-3498. 3

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, June 2020. ISSN: 2575-7075. 3

[20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 3

[21] C.-D. Nguyen, T. Nguyen, D. Vu, and A. Luu. Improving Multimodal Sentiment Analysis: Supervised Angular margin-based Contrastive Learning for Enhanced Fusion Representation. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14714–14724, Singapore, Dec. 2023. Association for Computational Linguistics. 3

[22] W. Yu, H. Xu, Z. Yuan, and J. Wu. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10790–10797, May 2021. Number: 12. 3, 7

[23] Y. Yu, M. Zhao, S.-a. Qi, F. Sun, B. Wang, W. Guo, X. Wang, L. Yang, and D. Niu. ConKI: Contrastive Knowledge Injection for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13610–13624, Toronto, Canada, 2023. Association for Computational Linguistics. 3

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 4

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4

[26] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6):82–88, Nov. 2016. 7

[27] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. 7

[28] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online, July 2020. Association for Computational Linguistics. 7

[29] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia, July 2018. Association for Computational Linguistics. 7