Dual Stream Alignment with Hierarchical Bottleneck Fusion For Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Multimodal sentiment analysis (MSA) leverages different modalities, such as text, image, and audio, for a comprehensive understanding of sentiment but faces challenges like temporal misalignment and modality heterogeneity. We propose a Dual-stream Alignment with Hierarchical Bottleneck Fusion (DAHB) method to address these issues. Our approach achieves comprehensive alignment through temporal alignment by cross-attention and semantic alignment via contrastive learning, ensuring alignment in time dimension and feature space. Moreover, Supervised contrastive learning is applied to refine these features. For modality fusion, we employ a hierarchical bottleneck method, progressively reducing bottleneck tokens to compress information and using bi-directional cross-attention to learn interactive between modalities. We conducted experiments on MOSI, MOSEI and CH-SIMS and results show that DAHB achieves state-of-the-art performance on a range of metrics. Ablation studies demonstrates the effectiveness of our methods. The code are available at url[1].

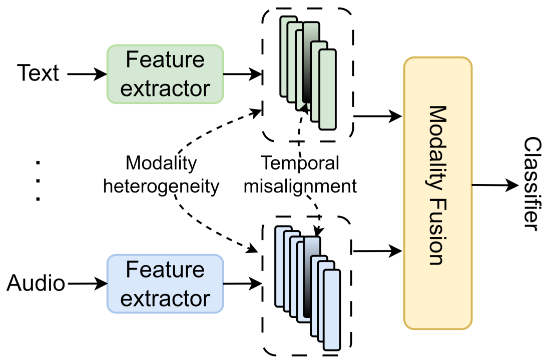

Figure 1: The temporal misalignment and modality heterogeneity in the pipeline of multimodal sentiment analysis (MSA).

1 Introduction

As an important component of human-computer interaction (HCI), sentiment analysis can enable computers to better understand and adapt to the emotional needs of humans (Wang et al., 2022). Compared to traditional text-based sentiment analysis, researchers have recently focused more on multimodal sentiment analysis (MSA), which involves using various data modalities (such as audio, text, and image) to infer and understand human emotional states. MSA leverages information from additional modalities, providing a more comprehensive view of sentiment. However, this also imposes significant challenges in effectively utilizing information from different modalities. The alignment and fusion of these diverse data sources are two of the primary challenges.

Alignment is the process of ensuring that information from different modalities is consistent in both time and semantic. As illustrated in Figure 1, MSA involves separating video into its components (text, image and audio) and independently extracting features from each. During this process, differences in sampling rates and preprocessing methods can cause features from different modalities at the same timestamp to not correspond correctly, leading to temporal misalignment that impairs accurate sentiment inference. However, misalignment exists not only in the time dimension but also in semantic due to the heterogeneity between different modalities. Each modality has distinct characteristics and representation space, which complicates seamless integration. Consequently, researchers (Li et al., 2021; Zong et al., 2023) have explored semantic alignment through contrastive learning, finding it can effectively enhance model performance. While some works have studied unilateral alignment, no research has simultaneously considered both temporal and semantic alignment.

Modal fusion, as the core component of MSA, aims to integrate complementary information from

[1] to ensure author anonymity, the link to the resource will be added after the review process

each modality. Current research has proposed various fusion mechanisms to achieve this integration. The two most common methods are directly utilizing cross-attention between different modality features and applying self-attention to the concatenation of unimodal features. Additionally, some studies (Lv et al., 2021; Sun et al., 2023) have introduced information hubs to facilitate communication between modalities. However, these methods include excessive redundant information, which can negatively impact effectiveness, and the quadratic computational complexity of attention mechanisms results in high computational costs.

Based on the above observations, we propose a dual-stream alignment with hierarchical bottleneck fusion (DAHB) framework. For multimodal data contains temporal information, we first utilize the dual-stream alignment to achieve comprehensive alignment in time and semantic space. For temporal alignment, we align audio and vision to the text in time dimension and obtain an well-aligned multimodal feature. For semantic alignment, features of different modalities from the same video are drawing closer in feature space, thus reducing the heterogeneity between modalities. After that, we introduce a supervised contrastive learning for both unimodal and multimodal features, to facilitate better feature discrimination and improve the model's robustness. Regarding modal fusion, inspired by Shwartz-Ziv and Tishby (2017), we leverage an attention bottleneck to integrate modalities similar to Nagrani et al. (2021) and achieve information compression by reducing the number of bottleneck tokens layer by layer. This progressive compression forces the model to learn the most beneficial sentiment representation. Our contributions can be summarized as follows:

- We propose a dual-stream alignment contains temporal alignment and semantic alignment, to realize the effective alignment between different modalities. Supervised contrastive learning is further introduced to improve the model's performance and robustness.

- We devise a novel hierarchical bottleneck fusion (HBF), which integrates different modality information through bottleneck and removing irrelevant information by compressing bottleneck layer by layer.

- We conduct comprehensive experiments on three publicly available datasets and gain superior or comparable results to the state-of-the-arts. Further studies verify the necessity of alignment and validity of our fusion mechanisms.

2 Related Work

In this section, we discuss the related work in MSA and contrastive learning.

2.1 Multimdoal Sentiment Anaylsis

Mainstream MSA approaches can be categorized into two types: fusion-based methods and representation learning-based methods.

Fusion-based methods primarily focus on designing sophisticated fusion mechanisms to obtain joint representations of multimodal data. Zadeh et al. (2017) used Tensor Fusion Networks (TFN) to obtain a tensor representation by computing the outer product of unimodal representations. Liu et al. (2018) designed a low-rank multimodal fusion method to reduce the computational complexity of tensor-based approaches. Tsai et al. (2019) proposed Cross-Modal Transformers, which learn cross-modal attention to enhance the target modality. Lv et al. (2021) introduced a message center to explore tri-modal interactions and perform progressive multimodal fusion. These methods perform fusion directly without considering the misalignment between the different modality features, which results in sub-optimal results.

Representation learning-based methods mainly focus on learning fine-grained modality semantics that encapsulate rich and diverse emotional cues, which can further enhance the effectiveness of multimodal fusion in relationship modeling. Hazarika et al. (2020) inspired by domain adaptation tasks, divided modality features into modality-invariant and modality-specific subspaces for multimodal fusion. Han et al. (2021) proposed MMIM, which improves multimodal fusion through hierarchical mutual information maximization. Guo et al. (2022) dynamically adjusted word representations in different non-verbal contexts using unaligned multimodal sequences. Nevertheless, these methods fail to considerate the impact of redundant information and fully exploit complementary information, which limits their performance in MSA.

2.2 Contrastive learning

Contrastive learning learns better data representation by drawing similar samples closer and pushing dissimilar samples further away in feature space.

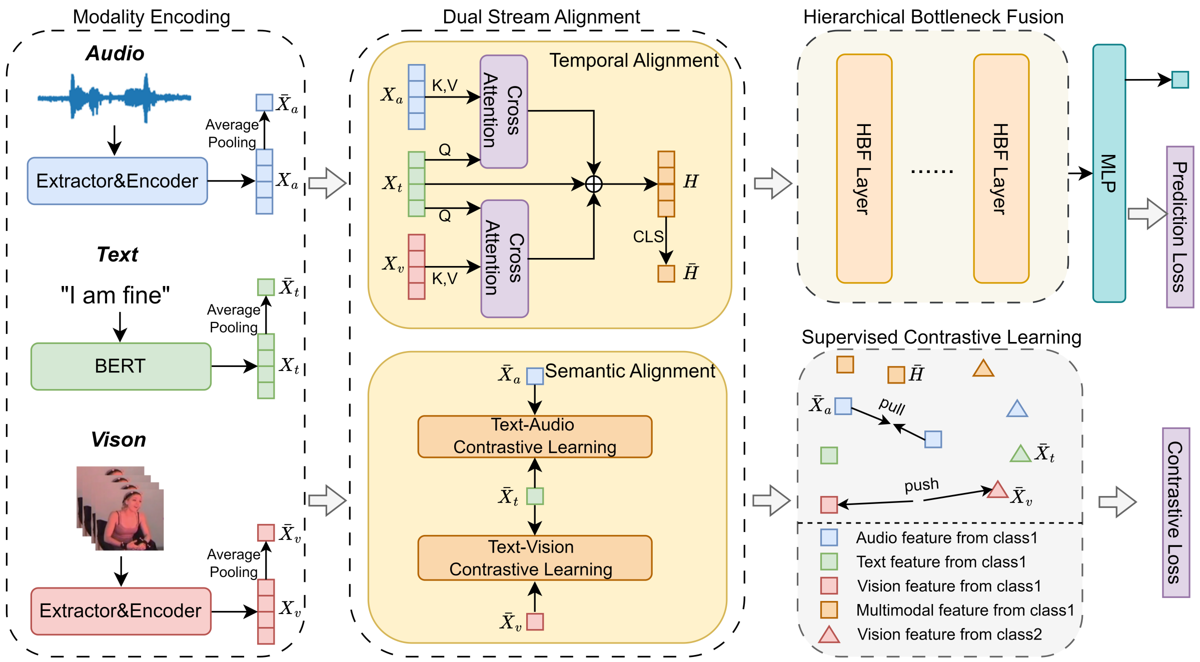

Figure 2: The overall architecture of the DAHB model for Multimodal Sentiment Analysis. It consists of modality encoding, dual-stream alignment, supervised contrastive learning, and hierarchical bottleneck fusion. Features are initially encoded independently, aligned temporally and semantically, refined through supervised contrastive learning, and finally fused via hierarchical bottleneck layers to produce robust sentiment predictions.

Since it does not require labels, contrastive learning has achieved significant success in self-supervised learning (Chen et al., 2020; He et al., 2020). Furthermore, because multimodal data inherent positive/negative pairs relations, contrastive learning has been widely applied in multimodal learning (Radford et al., 2021; Li et al., 2021). Khosla et al. (2020) extended contrastive learning to the supervised setting, they contrast samples by different classes and find it more stable for hyper-parameters. Recently, some MSA methods obtain modality representations based on contrastive learning. HyCon (Mai et al., 2022) simultaneously performed intra-/inter-modal contrastive learning to obtain tri-modal joint representations. Yang et al. (2023) decomposed each feature into similar and dissimilar parts for text-centered contrastive learning and designs a data sampler to retrieve positive/negative pairs. However, the existence of modality gap (Liang et al., 2022) makes it difficult to use contrastive learning alone to capture complementary information across different modalities.

3 Method

The overall architecture of DAHB is illustrated in Figure 2 . It consists of four parts: modality encoding, dual-stream alignment, supervised contrastive learning, and hierarchical bottleneck fusion. Our model first encodes each modality with corresponding feature extractors and encoders. Then, unimodal features are fed into the dual-stream alignment module to align in both the time dimension and feature space, producing aligned multimodal features. After that, the supervised contrastive learning module is employed to enhance the model's ability to distinguish different sentiment. Finally, we apply hierarchical fusion of modal features using the concept of information bottleneck. The fused unimodal features and multimodal features are concatenated and used to predict sentiment score. Below, we present the details of the four parts of DAHB.

3.1 Modality Encoding

Regarding the multimodal input, we first encode each modality into feature vectors. Following previous works (Yu et al., 2021; Han et al., 2021), we process raw audio and visual inputs into numerical sequential vectors using feature extractors (firmware with no parameters to train). Then, we employ two separate transformer encoders to encode these initial vector features. For the text modality, we use BERT to encode the text and scale it to the same feature dimension.

Then, we denote these modality features as

$X_m \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{t, v, a\}$, T_m is the sequence length and d_m is the vector dimension of each modality. In practice, T_m and d_m vary across different datasets.

3.2 Dual-Stream Alignment

We propose a dual-stream alignment method that includes both temporal and semantic alignment for comprehensive alignment. For temporal alignment, the unimodal features are dynamically aligned in the time dimension. For semantic alignment, we align matching modal pairs in the feature space. Furthermore, we choose text features as center in both temporal and semantic alignment, which can be viewed as connecting temporal and semantic alignment through text.

3.2.1 Temporal Alignment

In comparison to visual and speech signals, which are continuous and high-dimensional, text is discrete and contains more explicit semantic information. This discreteness and explicitness make text well-suited for alignment benchmark, as it allows for precise, word-by-word correspondence. Therefore, we align visual and speech modal features to text features.

Specifically, we use the Cross-Attention (CA) mechanism to achieve temporal alignment. CA can model the global dependencies relation of two modality sequences. The Query is from the target modality t, while the Key and Value are from the source modality s. In this way, CA can provide a latent adaptation from modality s to t :

$$\begin{aligned} \mathrm{CA}(X_t, X_s) &= \mathrm{softmax}\left(\frac{Q_t K_s^T}{\sqrt{d_k}}\right) V_s \\ &= \mathrm{softmax}\left(\frac{X_t W_Q W_K^T X_s^T}{\sqrt{d_k}}\right) X_s W_V \end{aligned} \tag{1}$$

where softmax represents weight normalization operation, W_Q and $W_K \in \mathbb{R}^{d \times d_k}$, $W_V \in \mathbb{R}^{d \times d_v}$ are learnable parameters and d_k is the dimension of attention head. Note that, for simplicity, we only present the formulation of single-head attention. In practice, we use multi-head CA (MHCA) to allow the model to attend to information from different feature subspaces.

In this way, we choose text features as Query, and speech features and vision features serve as the Key and Value, respectively. The aligned multimodal features H by aligning in time dimension are formalized as:

$$\begin{aligned} H &= V_t + V_{t \to a} + V_{t \to v} \\ &= V_t + \mathrm{CA}(X_t, X_v) + \mathrm{CA}(X_t, X_v) \end{aligned} \tag{2}$$

3.2.2 Semantic Alignment

Semantic alignment aims to draw close the features of matching modal pairs in feature space. Images, text, and audio from the same video are considered matching modal pairs. To achieve this, we utilize contrastive learning to align the semantic. This process maximizes a lower bound on the mutual information (MI) between different "views" of a video. Notably, because multimodal sentiment analysis remains largely centered around text information and to maintain consistency with feature-level alignment, we choose text as the anchor and the modality pairs are text-audio, text-vision. Specifically, we employ the NT-Xent loss (Chen et al., 2020) as the loss function for contrastive learning. The loss for sample i is defined as follows.

$$\ell_{\mathrm{cl}}^i = \sum_{(a,p) \in \mathcal{P}_i} -\log \frac{\exp(\mathrm{sim}(a,p)/\tau)}{\sum_{(a,k) \in \mathcal{N}_i \cup \mathcal{P}_i} \exp(\mathrm{sim}(a,k)/\tau)} \tag{3}$$

where τ is a temperature hyperparameter, (a, p), (a, k) correspond to the global features $\bar{X}_m$ of each modality, which are obtained by average pooling the unimodal features X_m along the time dimension. a represents the anchor in contrastive learning. $\mathcal{P}$ is the set of positive samples, and $\mathcal{N}$ is the set of negative samples. The similarity is measured by the dot product of the encoded archors and a set of encoded samples.

3.3 Supervised Contrastive Learning

To enhance the robustness of DAHB and fully utilize the information provided by the labels, we introduce supervised contrastive learning and unify it with semantic alignment in an NT-Xent loss framework. To maximize the potential of contrastive learning, we employ the hard negative mining approach. We construct positive and negative sample sets similar to the data sampler in Yang et al. (2023), which retrieves similar samples for a given sample based on both multimodal features and multimodal labels across samples. Note that, here we select not only unimodal features X_m for supervised contrastive learning but also multimodal features H obtained by temporal alignment.

First, we calculate the cosine similarity score between each sample pair (i, j) in dataset D. Then,

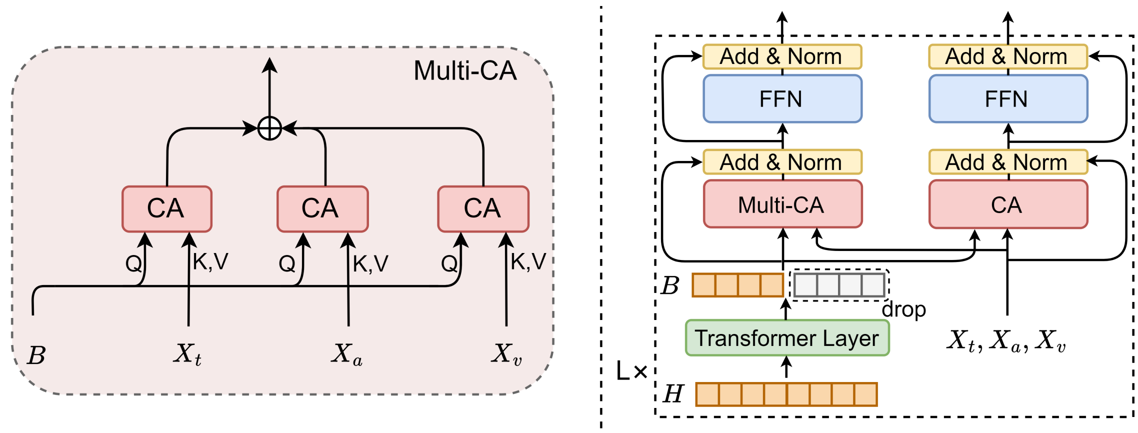

Figure 3: HBF Layer architecture. Multi-CA (left) gathers and integrate information from different modalitie through Cross-Attention.

we retrieve similar/dissimilar sample sets for each sample. For each sample i, we sort samples according to the similarity score. Two same class samples with high cosine similarity score are randomly selected to form positive pairs. For negative pairs, we randomly choose four samples with different labels: two that are similar to sample i and two that are dissimilar to sample i.

3.4 Hierarchical Bottleneck Fusion

For modality fusion, we use a bottleneck as a hub to facilitate communication with each modality features. At each layer, it reduces the number of bottleneck tokens and performs bidirectional cross-attention between the bottleneck and unimodal features. In this way, it allows the model to effectively integrate and compress multimodal information.

Specifically, the HBF layer is shown in Figure 3. We first introduce a Transformer Layer to encode the multimodal feature H and select the first p tokens as the bottleneck B. These tokens act as a compact summary of the multimodal information, capturing the most relevant features while discarding less important details. In each layer, the fusion is divided into two stages. Firstly, the bottleneck representation is used as Query to compute cross-attention with each of the three unimodal features (text, image and audio) and compress the refined multimodal information into the bottleneck representation. Secondly, each unimodal feature also performs cross-attention with the fused bottleneck representation B, updating the unimodal features. This step allows the unimodal features to incorporate information from other modalities. Additionally, the number of bottleneck tokens is halved in each layer. This progressive reduction process helps to further compress the information while preserving the essential features required for accurate sentiment analysis.

Suppose the HBF contains L layers. The overall equations of the l-th layer are formalized as.

$$B^l = \text{Transformer}(H^{l-1})[0 : p/2^{l-1}] \quad (4)$$

where B^l is bottleneck of the l-th layer.

$$\begin{aligned} H^l &= \text{LN}\left(B^l + \text{Multi-CA}(B^l, X_a^{l-1}, X_t^{l-1}, X_v^{l-1})\right) \\ H^l &= \text{LN}\left(H^l + \text{FFN}\left(H^l\right)\right) \end{aligned} \quad (5)$$

where LN denotes layer normalization, FFN is a feed-forward network with two linear transformations and a ReLU activation. H^l is multimodal features in the l-th layer.

$$\begin{aligned} Z_m^l &= \text{LN}\left(X_m^{l-1} + \text{CA}(X_m^{l-1}, B^l)\right) \\ X_m^l &= \text{LN}\left(Z_m^l + \text{FFN}\left(Z_m^l\right)\right), m \in \{t, v, a\} \end{aligned} \quad (6)$$

where $\mathbf{X}_m^l$ is unimodal features in the l-th layer.

3.5 Overall Learning Objectives

The DAHB model is trained with a multitask learning objective function, which consists of prediction loss and contrastive loss.

Prediction Loss. A multilayer perceptron (MLP) with ReLU activation function is used as a classifier to obtain the final prediction. We concatenate the first token of unimodal features and the bottleneck features after fusion to obtain the inputs to the classifier. The prediction loss is calculated by mean squared error.

$$\mathcal{L}_{pred} = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (7)$$

where n is the number of training samples and y_i is the sentiment label.

Contrastive Loss. As mentioned above, we unify the two modules of semantic alignment and supervised contrast learning through a simple joint contrastive loss. Specifically, this contrastive loss is expressed as:

$$\mathcal{L}_{con} = \frac{1}{n}\sum_{i=1}^{N}\ell_{\text{cl}}^{i} \quad (8)$$

where ℓ_{cl}^{i} is the contrastive loss of sample i.

Finally, the loss function of DAHB is represented as Equation (9), where λ is hyper-parameter to balance the contribution of each component to the overall loss.

$$\mathcal{L}_{all} = \mathcal{L}_{pred} + \lambda\mathcal{L}_{con} \quad (9)$$

4 Experiments

4.1 Datasets

We conduct experiments on three publicly available datasets in MSA research, MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020). The split specifications of the three datasets in Table 1. Here we give a brief introduction to the above datasets.

MOSI. As one of the most popular benchmark datasets for MSA, MOSI contains 2199 utterance-video clips sliced from 93 videos in which 89 distinct narrators are sharing opinions on interesting topics. Each clip is manually annotated with a sentiment value ranged from -3 (strongly negative) to +3 (strongly positive).

MOSEI. The dataset comprises 22,856 annotated video clips collected from YouTube. The MOSEI dataset upgrades MOSI by expanding the number of samples, utterances, speakers and topics. Its labeling style is same as MOSI.

CH-SIMS. The CH-SIMS dataset is a distinctive Chinese MSA dataset that contains 2,281 refined video clips collected from different movies, TV serials, and variety shows. Each samples has one multimodal label and three unimodal labels with a sentiment score from -1 (strongly negative) to 1 (strongly positive).

Table 1: Dataset statistics in MOSI, SOSEI, and SIMS.

Dataset	Train	Valid	Test	All
MOSI	1284	229	686	2199
MOSEI	16326	1871	4659	22856
CH-SIMS	1368	456	457	2281

4.2 Evaluation Metrics

Following previous works (Yu et al., 2021; Han et al., 2021; Yang et al., 2023), we report our results for classification and regression with the average of five runs of different seeds. For classification, we report the multi-class accuracy and weighted F1 score, i.e., 2-class accuracy (Acc-2), 3-class accuracy (Acc-3), and 5-class accuracy (Acc-5) and 7-class accuracy (Acc-7) for MOSI and MOSEI. Moreover, agreeing with prior works(Han et al., 2021; Yu et al., 2021), Acc-2 and F1-score on MOSI and MOSEI have two forms: negative/non-negative (non-exclude zero) and negative/positive (exclude zero). For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). Except for MAE, higher values indicate better performance for all metrics.

4.3 Baselines

To comprehensively validate the performance of our model, we compare our mothod with the several advanced and state-of-the-art baselines in Table 2 and 3: TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MulT (Tsai et al., 2019), MAG-BERT (Rahman et al., 2020), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), ConFEDE (Yang et al., 2023). To ensure fairness in comparison, the methods which only report the results of a single run and have no valid official code released for reproduction are not selected.

4.4 Performance Comparison

The performance comparison of all methods on MOSI, MOSEI, and CH-SIMS is summarized in Table 2 and Table 3.

As shown in Table 2, our method yields better or comparable results to many baseline methods, demonstrating the effectiveness of our approach in multimodal sentiment analysis (MSA). Specifically, on the MOSI dataset, our model outperforms all other baselines except for the Acc-7 metric. Additionally, our Acc-7 metric surpasses most of the baselines. For the MOSEI dataset, our model get

Table 2: Comparison on MOSI and MOSEI. † results from Yang et al. (2023), ‡ results from Han et al. (2021). All other results are reproduced using publicly available source codes and original hyper-parameters under the same setting. In Acc-2 and F1, the left of the "/" corresponds to "negative/non-negative" and the right corresponds to "negative/positive". (A) means the model utilized the aligned data.

Method	MOSI					MOSEI				
	Acc-2(↑)	F1(↑)	Acc-7(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	Acc-7(↑)	MAE(↓)	Corr(↑)
TFN†	-/80.8	-/80.7	34.9	0.901	0.698	-/82.5	-/82.1	50.2	0.593	0.700
LMF†	-/82.5	-/82.4	33.2	0.917	0.695	-/82.0	-/82.1	48.0	0.623	0.677
MulT(A)†	-/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(A)†	81.8/83.4	81.7/83.6	42.3	0.783	0.761	83.6/85.5	83.8/85.3	52.2	0.555	0.756
MAG-BERT†	82.13/83.54	81.12/83.58	41.43	0.790	0.766	79.86/83.86	80.47/83.88	50.41	0.583	0.741
Self-MM†	83.44/85.46	83.36/85.43	46.67	0.708	0.796	**83.76**/85.15	**83.82**/84.90	**53.87**	0.531	0.765
ConFEDE†	84.17/85.52	84.13/85.52	42.27	0.742	0.784	81.65/85.82	82.17/85.83	54.86	0.522	0.780
MMIM‡	84.14/86.06	84.0/85.98	46.65	0.70	0.800	82.24/85.97	82.66/85.94	54.24	0.526	0.772
ConFEDE	82.8/84.76	82.72/84.74	41.55	0.757	0.775	81.65/84.53	81.98/84.36	52.16	0.564	0.746
MMIM	83.46/85.11	83.4/85.24	**46.2**	0.714	0.794	81.64/85.24	81.84/85.19	53.23	0.538	0.763
Ours	**84.26/85.82**	**84.17/85.78**	45.63	**0.709**	**0.796**	82.27/**86.3**	82.7/**86.24**	53.12	**0.524**	**0.784**

Table 3: Comparison results on CH-SIMS. † results from Mao et al. (2022) and its corresponding GitHub page [1]. ‡ results from Yang et al. (2023). All other results are reproduced using publicly available source codes and original hyper-parameters under the same setting. (U) means the model used the multimodal label and unimodal label.

Method	CH-SIMS					
	Acc-2(↑)	F1(↑)	Acc-3(↑)	Acc-5(↑)	MAE(↓)	Corr(↑)
TFN†	78.38	78.62	65.12	39.30	0.432	0.591
LMF†	77.77	77.88	64.48	40.53	0.441	0.576
MulT†	78.56	79.66	64.77	37.94	0.453	0.561
MISA†	76.54	76.59	-	-	0.447	0.563
MAG-BERT†	74.44	71.75	-	-	0.492	0.399
self-MM†	80.04	80.44	65.47	41.53	0.425	0.595
ConFEDE(U)‡	82.23	82.08	70.15	46.30	0.392	0.637
Self-MM	78.56	78.60	64.68	41.69	0.428	0.585
ConFEDE(U)	81.1	80.95	68.93	45.43	0.387	0.643
Ours	**79.21**	**79.39**	**67.4**	**44.64**	**0.406**	**0.604**

best score in negative/positive (NP) setting for acc-2 and F1, MAE and Corr metrics. In particular, we have significant improvement on the NP Acc-2 and F1 score, indicating superior performance in distinguishing between positive and negative sentiments. For other metrics, our method also have comparable performance. However, in the negative/non-negative (NN) setting for Acc-2 and F1 metrics, our method does not perform as well as it does in the NP setting. This is because the NN setting is generally more challenging, requiring the model to classify data samples with a regression label of 0.

To further assess the effectiveness of our proposed method, DAHB, we conducted training on the CH-SIMS dataset. The scenarios in CH-SIMS are more intricate compared to those in MOSI and MOSEI, posing a greater challenge for modeling multimodal data. As seen in Table 3, for baselines that only use multimodal label, our method outperforms all of them on all metrics. Compared to the best baseline model, we achieve superior performance on multi-class, outperforming it by 2.4% on Acc-3 and 2.95% on Acc-5. Furthermore, our method performs closely to ConFEDE, which uses unimodal labels to enhance model training. Given that unimodal labels are difficult and time-consuming to obtain in real-world scenarios, our method demonstrates a significant advantage. These results highlight the robustness and practical applicability of DAHB in diverse and complex multimodal sentiment analysis tasks.

It is worth noting that Lian et al. (2024) found the MOSI and MOSEI datasets heavily emphasize the text modality, making it challenging for advanced fusion algorithms to showcase their advantages. In contrast, the CH-SIMS dataset is more balanced across modalities. Therefore, the CH-SIMS dataset provides a better platform to validate the integration of different modal information in our model, and we choose it for further ablation study.

4.5 Ablation Study and Analysis

4.5.1 Effects of Different Components

To evaluate the effectiveness of each component of our model, we conducted an ablation study by removing each component of DAHB individually. The results are shown in Table 4.

The experiment shows that all variations perform worse than the original model. Removing dual-stream alignment (the bottleneck is replaced by randomly initialized tokens) significantly decreases performance, which demonstrates that both temporal and semantic alignment positively impact the model's performance. Temporal alignment has

[1] https://github.com/thuiar/MMSA/blob/master/results/result-stat.md

a more conspicuous effect on multi-class accuracy, suggesting that aligning multimodal features along the time dimension provides more fine-grained sentiment information. On the other hand, omitting semantic alignment primarily affects two-class accuracy, highlighting the importance of aligning features from different modalities with the same semantic content.

Excluding supervised contrastive learning (SCL) results in a noticeable drop in performance, particularly in Acc-5, underscoring its role in enhancing the model's ability to effectively distinguish samples from different classes. The absence of hierarchical bottleneck fusion (HBF) leads to the most significant performance decrease, confirming its critical function in efficiently integrating and compressing multimodal information.

Table 4: The ablation study results on CH-SIMS.

Method	F1(↑)	Acc-5(↑)	MAE(↓)
DAHB	**79.39**	**44.64**	**0.406**
w/o dual-stream alignment	76.37	42.01	0.436
w/o teporal alignment	79.03	42.12	0.412
w/o semantic alignment	77.99	43.11	0.416
w/o SCL	78.65	41.79	0.420
w/o HBF	77.76	42.67	0.431

4.5.2 Effects of Different Fusion Mechanisms

To compare the effectiveness of different fusion mechanisms, we conducted experiments using various fusion mechanisms on the CH-SIMS dataset. The results, presented in Table 5, show the following observations.

The simplest method, concatenation, achieves moderate performance, indicating that when unimodal features are well-learned, simply combining them can be effective. However, it is not the most optimal approach for integrating multimodal information. Notably, this method incurs no additional multiply-accumulate operations (MAdds). Applying the self-attention (SA) mechanism to concatenated features significantly improves performance across all metrics, suggesting that self-attention enhances the learning of interactions among different modal features. However, this approach requires 324 million MAdds, indicating a substantial computational cost. The cross-attention (CA) mechanism integrates unimodal features into multimodal features and uses them for prediction. Although this method has a relatively low computational cost of 73 million MAdds, it performs poorly in terms of Acc-5 and MAE. This suggests that directly using cross-attention might lead to a loss of some feature details.

Bottleneck fusion (BF), which removing the compression process from our hierarchical bottleneck fusion, shows slightly better performance than simple concatenation. This demonstrates that using a bottleneck for fusion can help integrate multimodal features to some extent. Our proposed hierarchical bottleneck fusion (HBF) method achieves great improvement across most metrics and MAdds compared to bottleneck fusion. It delivers the best results in Acc-5 and MAE, confirming that the hierarchical approach of progressively reducing bottleneck tokens and using bi-directional cross-attention is highly effective in integrating and compressing multimodal information. Notably, the computational cost for our HBF is 145 million MAdds, which is less than half of that required by the self-attention mechanism (SA), demonstrateing that HBF can achieve superior performance while maintaining computational efficiency.

Table 5: Effects of different fusion mechanisms on CH-SIMS. The computation cost is measured by multiply-add operations (MAdds) with one video as the input. M denotes million.

Method	F1(↑)	Acc-5(↑)	MAE(↓)	MAdds
Concat	77.46	42.67	0.43	0
Concat&SA	**79.52**	44.38	0.414	324M
CA	78.53	39.95	0.456	73M
BF	78.56	42.89	0.453	162M
HBF	79.39	**44.64**	**0.406**	145M

5 Conclusion

In this paper, we propose a novel framework called DAHB aimed at enhancing multimodal sentiment analysis (MSA). To address temporal misalignment and heterogeneity across different modalities, we specifically design a dual-stream alignment mechanism consisting of temporal and semantic alignment. Additionally, we incorporate supervised contrastive learning to refine feature representations and enhance the model's robustness. Furthermore, we efficiently integrate modality features through hierarchical bottleneck fusion, employing bi-directional cross-attention for interaction and gradually reducing bottleneck tokens. Our methods achieve better performance than advanced methods on three prevalent datasets. Ablation studies and further analysis confirm the efficacy of our model and the necessity of each module.

Limitations

While our proposed DAHB method has demonstrated promising results in multimodal sentiment analysis, there are two limitations to consider. Firstly, although contrastive learning does not add extra parameters, the process requires significant GPU memory, necessitating more extensive sampling and training time. Moreover, the relatively small size of current sentiment analysis datasets introduces a level of randomness that may not accurately reflect the true performance of the model.

Acknowledgments

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

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

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3394–3402.

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

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Vecognition*, pages 9729–9738.

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. 2024. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *arXiv preprint arXiv:2401.03429*.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

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

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2554–2562.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

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: System Demonstrations*, pages 204–213, Dublin, Ireland. Association for Computational Linguistics.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213.

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

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52.

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

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, Online. Association for Computational Linguistics.

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

Amir Zadeh, 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, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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

A Baselines

TFN. The Tensor Fusion Network (Zadeh et al., 2017) calculates a multi-dimensional tensor utilizing outer product operations to capture uni-, bi-, and tri-modal interactions.

LMF. The Low-rank Multimodal Fusion (LMF)(Liu et al., 2018) decomposes stacked high-order tensors into many low rank factors to perform multimodal fusion efficiently.

MulT. The Multimodal Transformer (MulT)(Tsai et al., 2019) employs directional pairwise cross-modal attention to capture the interactions among multimodal sequences and adaptively align streams between different modalities.

MAG-BERT. The Multimodal Adaptation Gate for BERT (MAG-BERT)(Rahman et al., 2020) designs an alignment gate and insert that into different layers of the BERT backbone to refine the fusion process.

MISA. The Modality Invariant and -Specific Representations (MISA)(Hazarika et al., 2020) projects each modality features into modality-invariant and modality-specific spaces with special limitations. Fusion is then accomplished on these features.

SELF-MM. SELF-MM(Yu et al., 2021) assigns each modality a unimodal training task to obtain labels, then joint learn the multimodal and unimodal representations using multimodal and generated unimodal labels.

MMIM. MMIM(Han et al., 2021) proposes a hierarchical MI maximization framework that occurs at the input level and fusion level to reduce the loss of valuable task-related information.

HyCon. Hybrid Contrastive Learning of Tri-modal Representation (HyCon)(Mai et al., 2022) utilizes contrastive learning between modalities and classes to learn better modality representation.

ConFEDE. ConFEDE(Yang et al., 2023)is based on contrastive feature decomposition, which utilizes a unified contrastive training loss to capture the consistency and difference across modalities and samples.

B Experiments Setting

Here, we provide an overview of our experimental settings. All experiments were conducted on a single NVIDIA RTX 4090 GPU, with DAHB comprising fewer than 120 million parameters across all implementations.

For modality encoding, we use pretrained BERT models for text. Specifically, we employ "bert-base-chinese"[2] for CH-SIMS and "bert-base-uncased"[3] for MOSI and MOSEI. For vision and audio, we use transformers with 128 dimensions as Audio and Vision Encoders. For CH-SIMS and MOSI, we use two single-layer transformer encoders, while for MOSEI, we use three transformer layers due to its larger dataset size. In hierarchical bottleneck fusion (HBF), we set the number of bottleneck tokens, p, to 8. The number of fusion layers is set to 2 for MOSI and CH-SIMS, and 3 for MOSEI.

For model training, we train DAHB for MSA using the aforementioned encoders. The loss ratio, λ, is set to 0.2. For CH-SIMS and MOSI, we train DAHB for 100 epochs with a learning rate of 0.00005 and a batch size of 16. For MOSEI, we train the model for 25 epochs with a batch size of 8 and a learning rate of 0.00002.

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

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