AcFormer: An Aligned and Compact Transformer for Multimodal Sentiment Analysis

*Daoming Zong
SenseTime Group Limited
Beijing, China
ecnuzdm@gmail.com

*Chaoyue Ding
SenseTime Group Limited
Beijing, China
dingchaoyue@sensetime.com

Baoxiang Li
SenseTime Group Limited
Beijing, China
libaoxiang@sensetime.com

Jiakui Li
SenseTime Group Limited
Beijing, China
lijiakui@sensetime.com

Ken Zheng
SenseTime Group Limited
Beijing, China
zhengken@sensetime.com

Qunyan Zhou
SenseTime Group Limited
Beijing, China
zhouqunyan@senseauto.com

ABSTRACT

Multimodal Sentiment Analysis (MSA) is a popular research topic aimed at utilizing multimodal signals for understanding human emotions. The primary approach to solving this task is to develop complex fusion techniques. However, the heterogeneity and unaligned nature between modalities pose significant challenges to fusion. Additionally, existing methods lack consideration for the efficiency of modal fusion. To tackle these issues, we propose `AcFormer`, which contains two core ingredients: i) contrastive learning within and across modalities to *explicitly align* different modality streams before fusion; and ii) *pivot attention* for multimodal interaction/fusion. The former encourages positive triplets of image-audio-text to have similar representations in contrast to negative ones. The latter introduces attention pivots that can serve as cross-modal information bridges and limit cross-modal attention to a certain number of fusion pivot tokens. We evaluate `AcFormer` on multiple MSA tasks, including multimodal emotion recognition, humor detection, and sarcasm detection. Empirical evidence shows that `AcFormer` achieves the optimal performance with minimal computation cost compared to previous state-of-the-art methods. Our code is publicly available at https://github.com/dingchaoyue/AcFormer.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → *Multimedia information systems*; **Sentiment analysis**.

KEYWORDS

multimodal sentiment analysis; multimodal fusion; multimodal representation learning; contrastive learning; attention mechanism

*Daoming Zong and Chaoyue Ding contributed equally to this work.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611974

ACM Reference Format:
*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 (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611974

1 INTRODUCTION

With the prevalence of social networks, people are increasingly utilizing various media, such as text, images, and videos, to convey their emotions and opinions. Consequently, multimodal sentiment analysis (MSA) has emerged as a popular research area [2, 4, 52]. Most of the previous work on MSA has focused on developing complex fusion strategies, ranging from tensor-based [26, 55] fusion to attention-based [46] and MLP-based fusion [42], or subspace mapping schemes to extract commonalities among modalities and specificity of each modality [19, 54].

Despite achieving some successes, these approaches still have two shortcomings. Firstly, *the lack of explicit modality alignment before fusion.*Different modalities possess varying levels of noise topology and spatiotemporal redundancy, with some modality streams containing more task-relevant information than others [33]. For instance, visual and audio signals contain substantial spatiotemporal redundancy and useless noise information for other modalities due to the distinct sampling rates of each modality, making their features naturally unaligned [42]. Previous study also pointed out [22, 23] that learning the image-text interaction with a multimodal encoder becomes very challenging when image features and word embeddings are misaligned. Secondly, *the disregard of computational complexity in multimodal fusion.* Existing methods usually concatenate the representations of multiple single modalities and apply multi-head attention [49] to these representations, allowing each modality to attend to all the other modalities [19, 54]. Although applying full pairwise attention across all layers of a model is theoretically appealing, it is unnecessary because the visual and audio inputs contain abundant information, much of which is redundant. Besides, cross-modal attention is computationally intensive and time-consuming, with quadratic complexity when computing clip-to-clip correlations.

To address the aforementioned issues, we propose an aligned, compact Transformer, dubbed `AcFormer`, for multimodal sentiment

analysis. AcFormer has two attractive properties. Firstly, it reinforces the constraints on the representations of each modality prior to fusion through self-supervised learning [10, 20, 34], yielding more robust, compact, and well-aligned single-modality features. To attain the desired representations, we impose the intra-modal contrastive (IMC) and cross-modal alignment (CMA) losses on the outputs of unimodal encoders, with the aim of minimizing redundancy and noise in unimodal representations while promoting the later multimodal interactions. Specifically, CMA is designed for pulling together the embeddings of matched pairs (*i.e.*, image-text, image-audio, and text-audio) and pushing apart those of unmatched pairs, by maximizing the global mutual information of matched pairs. IMC attempts to maximize the agreement between different augmented views of the same data by maximizing their global mutual information. Secondly, it drastically reduces the computational cost of multimodal interactions while preserving the convergence/generalization rates of each modality, via the introduction of *pivot fusion tokens*. This permits free attention flow within each modality but restricts cross-modal attention flow, with each modality only able to interact with the tight *pivot fusion tokens*.

To sum up, our main contributions are as follows:

- We present AcFormer, an aligned and compact model for MSA. Experiments conducted on multiple benchmarks exhibit the superiority of AcFormer. Numerous ablation studies fully confirm the effectiveness of its components.
- We inject the intra-modal contrastive and cross-modal alignment loss on representations of the three unimodal encoders, which brings three benefits: (i) aligning the features from multiple modalities makes it easier for the fusion encoder to perform cross-modal learning; (ii) enhancing the unimodal encoder's understanding of intra-modal semantics; (iii) learning a shared low-dimensional space to embed images, audios, and texts, reducing their spatiotemporal redundancy and noise, resulting in more informative and compact unimodal representations.
- We restrict the cross-modal attention flow within each model layer through tight *pivot fusion tokens*. This compels the model to gather and compress information from each modality before sharing them, and only disseminate information most relevant to other modalities. By introducing a limited number of *pivot fusion tokens*, we drastically decrease the computational cost of cross-modal interaction.

2 RELATED WORK

2.1 Alignment before Fusion Matters

Previous work has shown that alignment before fusion facilitates multimodal representation learning [8, 17, 19, 46]. To begin with, GME-LSTM [8] posits the existence of local interactions among different modalities. To learn such local interactions, it employed the Penn Phonetics Lab Forced Aligner (P2FA) [21] to align words with their corresponding video and audio segments, where P2FA is a software which aligns an audio file and a verbatim text transcript. However, such word-level hard alignment is time-consuming. In addition to temporal alignment, an increasing number of works have focused on semantic alignment. For example, MulT [46] proposed an implicit alignment approach through directional pairwise cross-modal attention for multimodal human language time-series. MISA [19] and FDMER [54] reduce the semantic gap between modalities by mapping the representations of different modalities to a modality-invariant subspace and learning their commonality. CHFN [17] generates multimodal-shifted word representations by integrating visual and audio contexts using the self-attention mechanism. The aforementioned works all suggest that cross-modal alignment is beneficial for subsequent multimodal feature fusion.

2.2 Multimodal Representation Learning

Based on the granularity of fusion features, works on MSA can be roughly divided into two categories: *utterance-level* fusion and *word-level* fusion. Early works utilized features of the entire sentence, where utterance-level visual or acoustic features can be obtained by averaging frame-level visual/acoustic features. Utterance-level textual features can be obtained using sequence models such as LSTM or BERT [12]. Finally, the overall utterance-level features are fed into a fusion model to obtain a multimodal representation. Representative works in this line include multi-kernel learning [38] and tensor fusion (and its low-rank variants) [26, 30, 32, 55]. Utterance-level features mainly contain global information and may not capture local information. Therefore, recent works are more inclined towards word-level multimodal feature fusion. To extract word-aligned features, they first apply forced alignment to obtain timestamps for each word, from start time to end time. Then, the entire utterance is divided into several video segments according to the timestamps. Eventually, word-aligned visual or acoustic features are obtained by averaging the corresponding features of video segments. This line of methods covers word representation of non-verbal cues [17, 51], recurrent multistage fusion [24], graph fusion [31, 59], gated mechanism and LSTM fusion [8, 40, 56], MLP fusion [42] and a series of attention-based fusion models [46, 57].

3 APPROACH

Model Overview. Fig. 1 illustrates the overall architecture of AcFormer, which consists of a visual encoder $\mathbf{v}$, a text encoder $\mathbf{t}$, an audio encoder $\mathbf{a}$, and a set of attention bottleneck tokens for learning multimodal interactions. All of these encoders adopt the Transformer-style architecture [49], as detailed in Sec. 3.1. For each encoder, we maintain a momentum encoder, denoted by $\hat{\mathbf{v}}$, $\hat{\mathbf{t}}$, and $\hat{\mathbf{a}}$ and achieved via a momentum-based moving average strategy, following the same settings as in MoCo [20]. The unimodal encoders $\mathbf{v}$, $\mathbf{a}$, and $\mathbf{t}$ are responsible for extracting compact visual, acoustic and linguistic features from the given input, respectively. The pivot fusion tokens act on the deep layers of unimodal encoders, controlling the information flow between different modalities. Next, we will elaborate on each component in detail.

3.1 Unimodal Representation Learning

Given a triplet of video-audio-text ($\mathbf{V}, \mathbf{A}, \mathbf{T}$), we apply two separate augmentations to the video and audio to obtain two correlated *views*, namely (V_1, V_2) and (A_1, A_2), respectively. Akin to the vision transformer (ViT) [13], each augmentated video is divided into M video patches of $16\times16\times2$ by the video tokenizer composed of several 3D convolutions. For audio input, we use the speech augmentation policy [35] on the log-Mel spectrogram to obtain two

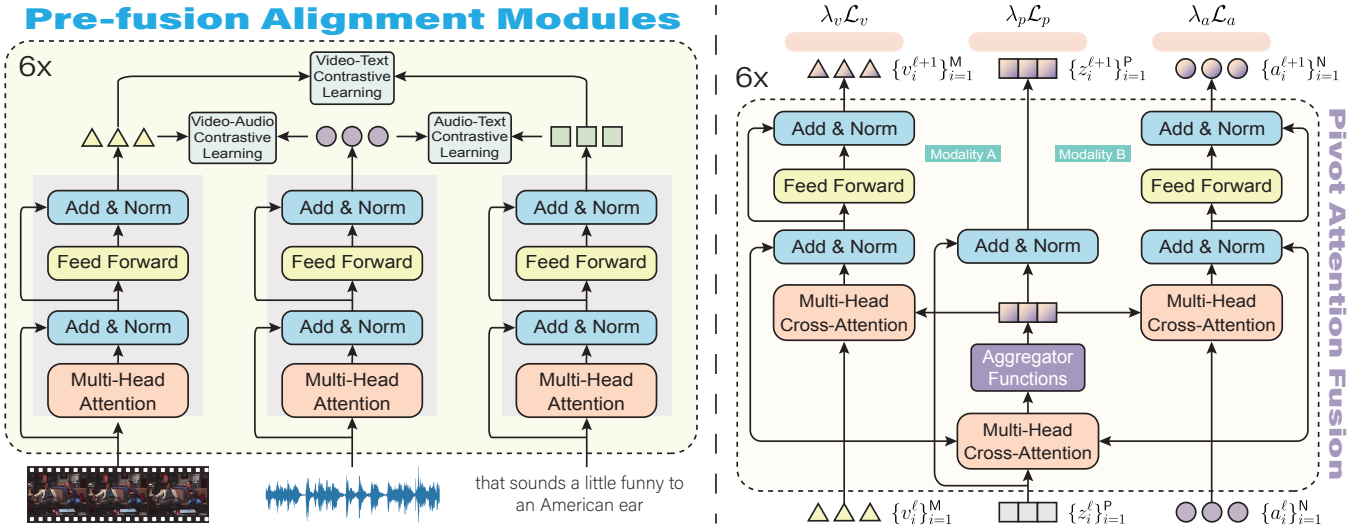

Figure 1: Illustration of AcFormer. It consists of a video encoder, an audio encoder and a text encoder. We design two pre-fusion alignment modules by leveraging intra-modal and cross-modal self-supervision to align unimodal representations of video-audio-text triplets before fusion. Furthermore, we also tailor a novel pivot attention fusion block for efficient multimodal interaction by bridging cross-modal communication via a small set of pivot fusion tokens $\{z_i\}$. AcFormer is trained using a weighted blending loss for each modality and the final pivot representations.

random views of the audio. Formally, let $\{v_1, \ldots, v_M\}$ be the desired representations of video $\mathbf{V}_1$ output by the visual encoder $\mathbf{v}$, and let $\{\hat{v}_1, \ldots, \hat{v}_M\}$ be the representations of video $\mathbf{V}_2$ output by the visual momentum encoder $\hat{\mathbf{v}}$. Likewise, denote by $\{a_1, \ldots, a_N\}$ the audio representations of $\mathbf{A}_1$, $\{\hat{a}_1, \ldots, \hat{a}_N\}$ the audio representations of $\mathbf{A}_2$, where N is the number of speech frames. For text input, we build the identical positive pairs by regarding dropout in BERT [12] as a minimal form of text augmentation [16]. We take BERT as a text encoder $\mathbf{t}$ and the momentum counterpart $\hat{\mathbf{t}}$, and derive $\{t_1, \ldots, t_L\}$ for $\mathbf{T}_1$ and $\{\hat{t}_1, \ldots, \hat{t}_L\}$ for $\mathbf{T}_2$, where $\mathbf{T}_1 = \mathbf{T}_2$, and L is the sequence length of text tokens. We found that appending [CLS] tokens to represent the semantics of the entire video/audio/text sequence results in suboptimal performance. Instead, we adopt the *token average pooling* to generate $\{v_{tap}, a_{tap}, t_{tap}\}$, which involves summing and averaging the representations of the last hidden states to better represent the unaligned multimodal sequence.

3.2 Cross-Modal Alignment Module (CMA)

The goal of CMA is to bring together the embeddings of matched modality pairs, while pushing apart those of unmatched modality pairs. These parallel modality pairs can be *video-text*, *video-audio*, or *audio-text* pairs. In other words, CMA aims to maximize the mutual information (MI) between two matched modalities, assuming that they describe the same semantics. Inspired by MoCo [20], we adopt the InfoNCE loss [34] as the contrastive learning loss function. Minimizing InfoNCE is equivalent to maximizing the lower bound of MI between the two modalities [34]. Formally, the InfoNCE loss for video-to-text is defined as follows:

$$\mathcal{L}^{v2t}(\mathbf{V}_1, \mathbf{T}_1, \tilde{\mathbf{T}}) = \mathbb{E}_{p(\mathbf{V},\mathbf{T})}\left[-\log \frac{\exp^{(sim(\mathbf{V}_1,\mathbf{T}_1)/\tau)}}{\sum_{i=1}^{K}\exp^{(sim(\mathbf{V}_1,\tilde{\mathbf{T}}_k)/\tau)}}\right], \quad (1)$$

where τ is a temperature hyper-parameter, and $\tilde{\mathbf{T}} = \{\tilde{\mathbf{T}}_1, \ldots, \tilde{\mathbf{T}}_K\}$ is a set of negative text samples that are not matched to $\mathbf{V}_1$. The similarity is measured by dot product on an encoded query and a set of encoded samples, *i.e.*, $sim(\mathbf{V}_1, \mathbf{T}_1) = \mathbf{q}_v(v_{tap})^\top \hat{\mathbf{k}}_t(t_{tap})$, where $\mathbf{q}_v$ and $\hat{\mathbf{k}}_t$ are two projection heads to generate *key* and *query*. To yield the negative text samples $\tilde{\mathbf{T}}$, we maintain a large queue to store the most recent K projected representations $\hat{\mathbf{k}}_t(t_{tap})$, as in [20]. The InfoNCE loss for text-to-video can be built in a similar way:

$$\mathcal{L}^{t2v}(\mathbf{T}_1, \mathbf{V}_2, \tilde{\mathbf{V}}) = \mathbb{E}_{p(\mathbf{T},\mathbf{V})}\left[-\log \frac{\exp^{(sim(\mathbf{T}_1,\mathbf{V}_2)/\tau)}}{\sum_{i=1}^{K}\exp^{(sim(\mathbf{T}_1,\tilde{\mathbf{V}}_k)/\tau)}}\right], \quad (2)$$

where $\tilde{\mathbf{V}} = \{\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_K\}$ is a set of negative image samples that are not matched to $\mathbf{T}_1$. And $sim(\mathbf{T}_1, \mathbf{I}_2) = \mathbf{q}_t(t_{tap})^\top \hat{\mathbf{k}}_v(\hat{v}_{tap})$, where $\mathbf{q}_t$ and $\hat{\mathbf{k}}_v$ are two projection heads. Similarly, we also maintain a queue to store the most recent K projected representations $\hat{\mathbf{k}}_v(\hat{v}_{tap})$.

We also consider the InfoNCE loss for (video $\rightleftharpoons$ audio) and (text $\rightleftharpoons$ audio) in both directions, denoted by ($\mathcal{L}^{v2a}$,$\mathcal{L}^{a2v}$) and ($\mathcal{L}^{t2a}$, $\mathcal{L}^{a2t}$), respectively. Their definitions are very similar to that of (video $\rightleftharpoons$ text). Overall, the optimization objective of CMA is:

$$\mathcal{L}_{cma} = \frac{1}{6}[(\mathcal{L}^{v2t} + \mathcal{L}^{t2v}) + (\mathcal{L}^{v2a} + \mathcal{L}^{a2v}) + (\mathcal{L}^{t2a} + \mathcal{L}^{a2t})]. \quad (3)$$

Remarks. Theoretically, by minimizing $\mathcal{L}_{cma}$, we encourage visual, acoustic and linguistic features to align well in the embedding space, thus simplifying the fusion of multimodal features. However, the correspondence between multiple modalities is not always perfect and may contain a significant amount of noise.

3.3 Intra-Modal Contrastive Learning (IMC)

IMC aims to learn generalized representations by maximizing agreement between differently augmented views of the same data sample via a contrastive loss [10]. As stated in Sec. 3.1, for visual input, two random views ($\mathbf{I}_1$, $\mathbf{I}_2$) of the same image $\mathbf{I}$ under image data augmentation are considered positive pairs. Likewise, two random views ($\mathbf{A}_1$, $\mathbf{A}_2$) of the same audio $\mathbf{A}$ are regarded as positive pairs. As text data augmentation is inherently challenging due to its discreteness,

we construct positive pairs by setting $\mathbf{T}_1 = \mathbf{T}_2$ as suggested by [16], *i.e.*, predicting the input text itself as the contrastive object. This is achieved by treating standard dropout as the minimal data augmentation for text and independently applying sampled dropout masks to $\mathbf{T}_1$ and $\mathbf{T}_2$. Similar to CMA, we construct a dynamic dictionary with a queue for each modality and employ InfoNCE loss. Formally, the optimization objective of IMC can be defined as:

$$\mathcal{L}_{imc} = \frac{1}{3}\left[\mathcal{L}^{v2v}(\mathbf{V}_1, \mathbf{V}_2, \tilde{\mathbf{V}}) + \mathcal{L}^{a2a}(\mathbf{A}_1, \mathbf{A}_2, \tilde{\mathbf{A}}) + \mathcal{L}^{t2t}(\mathbf{T}_1, \mathbf{T}_2, \tilde{\mathbf{T}})\right]. \tag{4}$$

Remarks. Two key properties related to $\mathcal{L}_{imc}$ should be noted here, namely (i) *alignment* (closeness) of embeddings from semantically-related positive pairs, and (ii) *uniformity* of the induced distribution of the (normalized) embeddings on the hypersphere [16, 50]. Prior study [50] has proven that the contrastive loss asymptotically optimizes the two properties. Therefore, optimizing $\mathcal{L}_{imc}$ encourages the formation of positive pairs of samples with the smallest distance in the embedding space, while keeping robust to noise disturbances. Meanwhile, the embeddings should be evenly distributed on the unit hypersphere, preserving as much informative data as possible.

3.4 Video-Audio-Text Matching (VAT-M)

VAT-M predicts whether a triplet of ⟨video, speech, text⟩ is positive (matched) or negative (non-matched). The output tokens of each modality in the last layers, *i.e.* $\{v_{tap}, a_{tap}, t_{tap}\}$, are summed and averaged to serve as the joint representation of the triplet, which are then fed into a fully connected layer (FC) to predict the binary probability. The VAT-M loss is written as:

$$\mathcal{L}_{vat-m} = \mathbb{E}_{(\mathrm{V,A,T})\sim\mathcal{D}}\mathrm{H}(p(\mathrm{I,A,T}), y^{(\mathrm{V,A,T})}), \tag{5}$$

where $p(\cdot)$ denote the output probability, $y^{(\mathrm{V,A,T})}$ is the ground-truth label. H here denotes the binary cross-entropy loss.

3.5 Modality Fusion Via Pivot Attention

To capture cross-modal global correlations, one commonly used solution in MSA is to concatenate the representations of all single modalities [19, 54], and then utilize self-attention [49] to learn modality interactions. However, the limitation of such strategy lies in the high computational cost of self-attention, which has a quadratic complexity in calculating clip-to-clip correlations. Recently, MBT [33], PMR [29], UMT [25] and EMT [43] leverage the concept of *message hub* to communicate with each modality. The message hub can propagate common messages to each modality and it can also collect useful information from them. We also extend the core idea of attention bottlenecks for multimodal feature fusion by decoupling it into: *feature aggregation* and *feature scattering*.

3.5.1 Feature Aggregation. The aim of feature aggregation is to *gather* and *condense* information from different modalities. Similar to [33], we introduce a set of extra pivot tokens $\mathbf{Z} = \{z_i\}_{i=1}^{p}$ to connect with all modalities, which is achieved by multiple multi-head attentions between features from different modalities. Note that p is a much smaller number than that of video patches, audio frames, or text tokens. Take one *head* as an example, the process of feature aggregation can be formulated by:

$$\mathbf{Z}^{(\ell+1)} = \mathbf{Z}^{(\ell)} + \mathrm{AGGREGATE}\Big[\ \mathrm{softmax}(\frac{QK_m^\top}{\sqrt{d}})V_m\Big],\ m \in \{v, a, t\}, \tag{6}$$

where $\mathbf{Z}^{(\ell)}$ and $\mathbf{Z}^{(\ell+1)}$ are the input and output features of pivot tokens after a transformer block. $Q = \mathbf{Z}W^q$ denotes the shared *query* matrices across all modalities, where W^q is the learnable parameter matrix. $K_{\{v,a,t\}}$ and $V_{\{v,a,t\}}$ are the modality-specific *key*, *value* matrices, yielded by linear projections of each modality features. We compute attention scores separately for each modality and compress refined multimodal information into pivot token representations via aggregation function and residual connection. A wide range of aggregation functions are applicable, such as *summation*, *mean*, *max pooling* and *concat*. We found that different aggregators only result in minor performance differences, as evidenced in Sec. 4.5.

3.5.2 Feature Scattering. The purpose of feature scattering is to propagate the aggregated features across distinct modalities, enabling each modality to perceive information from other modalities. The information delivered to each modality is controlled by another multi-head attention [49], proceeding as follows:

$$\begin{aligned}\mathbf{X}_m^{(\ell)} &= \mathrm{LayerNorm}(\mathbf{X}_m^{(\ell)} + \mathrm{softmax}(\frac{Q_m K_z^\top}{\sqrt{d}})V_z),\\ \mathbf{X}_m^{(\ell+1)} &= \mathrm{LayerNorm}(\mathbf{X}_m^{(\ell)} + \mathrm{FFN}(\mathbf{X}_m^{(\ell)})),\ m \in \{v, a, t\},\end{aligned} \tag{7}$$

where $\mathbf{X}_{\{v,a,t\}}^{(\ell)}$ are the learned features in the l-th layer of visual, audio, and text encoder, respectively. Q_v, Q_a, and Q_t are modality-specific *query* matrices, generated by $\mathbf{X}_v W_v^q$, $\mathbf{X}_a W_a^q$, and $\mathbf{X}_t W_t^q$ respectively, where $W_{\{v,a,t\}}^{(q)}$ are learnable weight matrices. K_z, V_z are shared *key*, *value* matrices produced by the linear projections of pivot tokens, *i.e.* $\mathbf{Z}W^{(k)}$ and $\mathbf{Z}W^{(v)}$. FFN is a feed-forward network with two linear transformations and a ReLU activation.

3.6 Optimization Object

The training of AcFormer consists of two phases. In Phase I, the first 6 layers of all unimodal encoders are pre-trained on a given dataset, using the following optimization objective:

$$\mathcal{L}_{rep} = \mathcal{L}_{cma} + \mathcal{L}_{imc} + \mathcal{L}_{vat-m}. \tag{8}$$

In Phase II, the first 6 encoder layers are fine-tuned on task-related datasets, and then the pivot attention tokens are trained along with the entire AcFormer via the task-specific loss. Specifically, for the classification task, we adopt the standard cross-entropy loss as $\mathcal{L}_{task} = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log \hat{y}_i$, while for the regression task, we use the mean squared error loss as $\mathcal{L}_{task} = -\frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2$, where y_i is the ground truth and n is the number of samples.

4 EXPERIMENTS

4.1 Datasets

We evaluate AcFormer on three commonly-used multimodal emotion recognition (MER) benchmarks, a multimodal sarcasm detection (MSD) and a multimodal humor detection (MHD) benchmarks. Due to space limitations, we defer the introduction of these widely-used datasets, including **CMU-MOSI** [58], **CMU-MOSEI** [59], **IEMOCAP** [5], **UR-FUNNY** [18] and **MUStARD** [6] to Appendix A.

Apart from the tasks of MSD and MHD, we evaluated all the MER tasks in two different settings: *word-aligned setting* and *unaligned setting*. Specifically, the word-aligned setting relies on an additional step to manually align the visual and audito streams at the word-level resolution in the text to obtain word-level aligned multimodal features. We trained AcFormer♠ using the same word-aligned features as in the previous methods [19, 42, 54]. In the unaligned setting, we used our own extracted raw multimodal sequences to train AcFormer♣. For a fair comparison, we only used pre-trained weights in the unaligned setting. We employed the Multimodal EmotionLines Dataset (**MELD**) as the pre-training dataset.

MELD [39] is an enhanced and extended version of the EmotionLines [9] dataset. MELD comprises data from three modalities, namely visual, audio, and text, with over 1400 conversations and a total of 13000 utterances from the popular TV show "Friends". Multiple speakers are involved in these conversations, and each utterance is labeled with one of the seven emotions: anger, disgust, sadness, joy, neutral, surprise, and fear. Additionally, MELD provides annotations for the emotional polarity of each utterance.

4.2 Feature Extraction

To ensure a fair comparison, we utilize the same low-level embeddings as previous works [19] for the word-aligned settings. We also elaborate on the raw multimodal feature extraction process for the non-aligned settings.

Textual Features. Under **word-aligned settings**, we first consider the pre-trained 300-dimensional Glove embeddings [36] for each token in an utterance, as done in many preivous works [6, 19, 47]. We also utilize the pre-trained BERT-base-uncased [12] to extract a sequence of 768-dimensional token representations. Unlike prior works that opt for a unique utterance representation by either averaging the last hidden layer outputs of BERT or using the special [CLS] token [6, 19], we directly take the token-wise representations output by pre-trained BERT. We report results using both Glove and BERT for a comprehensive and fair comparison. For **unaligned settings**, we also extract 300-dimensional pre-trained Glove word embeddings and 768-dimensional BERT-base-uncased token embeddings of the last hidden state for each video transcript.

Visual Features. Under the **word-aligned setup**, we adopted the same word-level aligned visual features as in [6, 19, 46, 47]. Concretely, for MOSI and MOSEI, facial expression features including facial action units and poses based on the Facial Action Coding System (FACS) [41], are extracted using Facet. For UR-FUNNY, features related to facial expressions of speakers are extracted using OpenFace [3], a facial behavior analysis toolk. For IEMOCAP, video frames are processed by Facet to generate facial action units and poses that represent the facial muscle movement. This leads to a sequence of frame-level visual representations, with the feature dimensions of 47 for MOSI, 35 for MOSEI, 75 for UR-FUNNY and 35 for IEMOCAP. For **unaligned settings**, given a video clip $\mathbf{x}_v$ of t seconds duration, we crop and align faces for each video frame using OpenFace and uniformly sample 8 frames $\mathbf{x}_v \in \mathbb{R}^{\mathrm{c}\times 8\times \mathrm{h}\times \mathrm{w}}$. Then we use a video tokenizer, consisting of several 3D convolution and max pooling layers with frame kernel size of 2 and kernel size of 16, to convert the 8 sampled frames into a total of 196 video patches $\mathbf{x}_v \in \mathbb{R}^{196\times 768}$, each with a feature dimension of 768.

Acoustic Features. For **word-aligned settings**, following the previous works [6, 19, 46, 47], we employed various low-level statistical audio functions extracted from an acoustic analysis toolkit, COVAREP [11], including Mel-frequency cepstral coefficients (MFCCs), pitch, voiced/unvoiced segmentation features [14], glottal source parameters [15], and other emotion- and prosody-related features. This results in a feature dimension of 74 for MOSI/MOSEI, 81 for UR-FUNNY and 74 for IEMOCAP. For **non-aligned settings**, given a raw audio of t seconds length, we extract log Mel-filterbank (fbank) features with number of mel-frequency bins of 128 computed with a 25ms Hamming window every 10ms. This produces a sequence of audio tokens $\mathbf{x}_a \in \mathbb{R}^{100t\times 128}$ to AcFormer♣.

4.3 Implementation Details

We train two variants of AcFormer for extensive evaluation. For the word-aligned setting, we train AcFormer♠ equipped with a 5-layer unimodal encoder per modality and a 2-layer fusion encoder. And we do not use pre-trained weights in such settings. Both the unimodal encoder and the fusion encoder have the same embedding dimension, which is set to 768. Before feeding the input of each modality into the unimodal encoders, three separate linear layers are applied to map these inputs to the same dimension. Unless otherwise specialized, we set the number of pivot tokens to 12 and initialize them using a Gaussian with mean of 0 and standard deviation of 0.02. For the unaligned setting, we first pre-train AcFormer♣ on MELD and fine-tune it on each task-specific dataset. AcFormer♣ contains a 12-layer vision encoder $\mathbf{v}$, a 12-layer audio encoder $\mathbf{a}$, a 12-layer text encoder $\mathbf{t}$, and 12 learnable pivot tokens $\mathbf{z}_{fsn}$. For the visual encoder, we use a 12-layer Vision Transformer ViT-B/16 [13] and initialize it with weights pre-trained on ImageNet-1k from [45]. All sampled frames from a given video clip $\mathbf{V}$ are first processed by a video tokenizer to generate a sequence of video tokens. Concretely, we initialize the text encoder using the pre-trained weights from BERT-base [12] and initialize the audio encoder with the pre-trained weights from wav2vec2.0 [1].

Pre-training. As illustrated in Fig. 1, we only pre-trained the first 6 layers of visual, audio and text encoders. We pre-trained AcFormer♣ with three objectives: cross-modal alignment (CMA), intra-modal contrastive learning (IMC) as well as the video-audio-text matching (VAT-M) on the unimodal encoders. All our experiments were conducted on eight NVIDIA V100 GPUs, each with 32GB of memory. In the pre-training stage, AcFormer♣ was trained for 60 epochs with a batch size of 4 per GPU, taking approximately 6 hours. We have trained two version of models, one based on GloVe word embeddings, *i.e.* AcFormer♣ (G), and the other based on BERT token embeddings, *i.e.* AcFormer♣ (B). We use the AdamW [27] optimizer with a weight decay of 0.02. The learning rate is initialized as $1e{-}5$ and is warmed up to $1e$–4 after 10 training epochs. We then decrease it by the cosine decay strategy to $1e{-}5$. All sampled video frames are resized to the same resolution of 224×224 and some common image augmentations are applied, such as random color jittering, random grayscale conversion and random Gaussian Blur, etc.

Fine-tuning. During fine-tuning, we first jointly optimize the combined loss of CMA, IMC and VAT-M on task-specific datasets. Next, we removed all contrastive projection heads used in the pre-training

Models	CMU-MOSI				
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F-Score (↑)	Acc-7 (↑)
TFN [55]	0.970	0.633	73.9 / -	73.4 / -	32.1
MARN [57]	0.968	0.625	77.1 / -	77.0 / -	34.7
MFN [56]	0.965	0.632	77.4 / -	77.3 / -	34.1
LMF [26]	0.912	0.668	76.4 / -	75.7 / -	32.8
MFM [47]	0.951	0.662	78.1 / -	78.1 / -	36.2
RAVEN [51]	0.915	0.691	78.0 / -	76.6 / -	33.2
MCTN [37]	0.909	0.676	79.3 / -	79.1 / -	35.6
CIA [7]	0.914	0.689	79.8 / -	- / 79.5	38.9
MulT [46]	0.871	0.698	- / 83.0	- / 82.8	40.0
TCSP [53]	0.908	0.710	- / 80.9	- / 81.0	-
PMR [29]	-	-	- / 83.6	- / 83.4	40.6
PDMER [54]	0.845	0.732	- / 84.2	- / 83.9	42.1
AcFormer♠	0.833	0.741	80.7/85.2	80.8/84.6	42.7
AcFormer♣	0.796	0.806	82.1/86.3	82.0/85.9	43.8
TFN (B) [55]	0.901	0.698	- / 80.8	- / 80.7	34.9
LMF (B) [26]	0.917	0.695	- / 82.5	- / 82.4	33.2
MFM (B) [47]	0.877	0.706	- / 81.7	- / 81.6	35.4
ICCN (B) [44]	0.860	0.710	- / 83.0	- / 83.0	39.0
CubeMLP (B) [42]	0.770	0.767	- / 85.6	- / 85.5	45.5
MISA (B) [19]	0.783	0.761	81.8 / 83.4	81.7 / 83.6	42.3
FDMER (B) [54]	0.724	0.788	- / 84.6	- / 84.7	44.1
AcFormer♠ (B)	0.715	0.794	82.3/85.4	82.1/85.2	44.2
AcFormer♣ (B)	0.703	0.815	83.1/86.4	83.2/86.7	45.3

Table 1: Performances of multimodal models in MOSI. NOTE: (B) means the language features are based on BERT.

Models	CMU-MOSEI				
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F-Score (↑)	Acc-7 (↑)
MFN$^{\otimes}$ [56]	-	-	76.0 / -	76.0 / -	-
Graph-MFN$^{\otimes}$ [58]	0.710	0.540	76.9 / -	77.0 / -	45.0
RAVEN [51]	0.614	0.662	79.1 / -	79.5 / -	50.0
MCTN [37]	0.609	0.670	79.8 / -	80.6 / -	49.6
CIA [7]	0.680	0.590	80.4 / -	78.2 / -	50.1
MulT [46]	0.580	0.703	- / 82.5	- / 82.3	51.8
TCSP [53]	0.576	0.715	82.8	82.7	-
PMR [29]	-	-	83.3	82.6	52.5
FDMER [54]	0.568	0.736	83.9	83.8	53.8
AcFormer♠	0.556	0.742	81.1/84.5	82.0/83.8	54.0
AcFormer♣	0.532	0.772	83.6/85.7	83.8/85.9	54.3
TFN (B)$^{\diamond}$ [55]	0.593	0.700	- / 82.5	- / 82.1	50.2
LMF (B)$^{\diamond}$ [26]	0.623	0.677	- / 82.0	- / 82.1	48.0
MFM (B)$^{\diamond}$ [47]	0.568	0.717	- / 84.4	- / 84.3	51.3
ICCN (B) [44]	0.565	0.713	- / 84.2	- / 84.2	51.6
MISA (B) [19]	0.555	0.756	83.6 / 85.5	83.8 / 85.3	52.2
CubeMLP (B) [42]	0.529	0.760	- / 85.1	- / 84.5	54.9
FDMER (B) [54]	0.536	0.773	- / 86.1	- / 85.8	54.1
AcFormer♠ (B)	0.531	0.786	84.3/86.5	84.2/85.8	54.7
AcFormer♣ (B)	0.518	0.802	86.1/88.4	86.3/88.0	56.9

Table 2: Performances of multimodal models in MOSEI. NOTE: (B) means the language features are based on BERT.

Setting	Algorithms	Happy		Sad		Angry		Neutral	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Aligned	MFM [47]	90.2	85.8	88.4	86.1	87.5	86.7	72.1	68.1
	RAVEN [51]	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3
	MCTN [37]	84.9	83.1	80.5	79.6	79.7	80.4	62.3	57.0
	MulT [46]	90.7	88.6	86.7	86.0	87.4	87.0	72.4	70.7
	PMR [29]	91.3	89.2	87.8	87.0	88.1	87.5	73.0	71.5
	AcFormer♠	**91.3**	**89.8**	**88.6**	**87.3**	**88.4**	**87.8**	**73.7**	**72.0**
	AcFormer♠ (B)	**92.7**	**91.3**	**90.1**	**89.5**	**89.9**	**88.2**	**75.4**	**73.2**
Unaligned	RAVEN [51]	77.0	76.8	67.6	65.6	65.0	64.1	62.0	59.5
	MCTN [37]	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3
	MulT [46]	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
	PMR [29]	86.4	83.3	78.5	75.3	75.0	71.3	63.7	60.9
	AcFormer♣	**90.6**	**89.3**	**89.7**	**88.4**	**87.9**	**86.2**	**72.3**	**71.1**
	AcFormer♣ (B)	**92.4**	**91.1**	**91.7**	**90.3**	**89.8**	**88.1**	**74.0**	**72.6**

Table 3: Performance Comparison (%) on IEMOCAP under both the word-aligned setting and the unaligned setting.

stage, and assembled the unimodal encoders with the last six transformer layers and the associated pivot tokens to form the complete AcFormer♣. Finally, we retrain the entire AcFormer♣ with a small learning rate using the task-specific loss. For regression tasks, we use the *mean square error* loss whereas for classification tasks, we use the *cross-entropy* loss. Since different modalities have various learning dynamics generalize at different rates, training them jointly with a single optimization goal may be sub-optimal. To solve this, we joint train three modalities using the weighted blending loss of $\lambda_v\mathcal{L}_v, \lambda_a\mathcal{L}_a, \lambda_t\mathcal{L}_t$, and $\lambda_p\mathcal{L}_p$. The hyperparameters $\{\lambda_v, \lambda_a, \lambda_t, \lambda_p\}$ are empirically set to $\{0.2, 0.2, 0.4, 0.2\}$. See Supp. B for more details.

4.4 Main Results

Multimodal Emotion Recognition. Tables 1, 2, and 3 report performance comparisons of different baseline methods on the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets, respectively. Overall, AcFormer outperforms previous state-of-the-art methods in all metrics across all benchmarks. Notably, AcFormer surpasses some methods with sophisticated modality fusion mechanisms, such as TFN, LFN, and PMR. This indicates that our modality fusion scheme on top of pivot attention tokens can learn sound multimodal representations. Additionally, we observe that subspace mapping methods, such as MISA and PDMER, the former uses orthogonal constraints to learn modality-invariant and -specific subspaces while the latter uses GANs to learn modality-invariant and -specific subspaces, perform competitively. This confirms the importance of learning modality representations before fusion. Nevertheless, our approach surpasses these competitors, demonstrating the effectiveness of our use of contrastive learning to align inter-modal and intra-modal representations.

Multimodal Humor/Sarcasm Detection. Understanding humor requires the exploration of intra-modal information and inter-modal interactions among multiple modalities. Identifying sarcasm also requires additional cues beyond the text, such as the incongruity between modalities or the concomitant emotions expressed within modalities [6]. We conduct experiments on UR-FUNNY and MUStARD benchmarks to evaluate the generalization ability of AcFormer. The results (see Supp. C and D) show that AcFormer achieves the best performance in both multimodal humor detection and sarcasm detection tasks, indicating that AcFormer indeed acquires a more discriminative multimodal representation that enhances its prediction of humor or sarcasm.

4.5 Ablation Study and Analysis

In this section, we meticulously examine the efficacy of main components, as well as the computational cost of AcFormer.

4.5.1 ***Impact of Pre-fusion Modality Alignment.*** As shown in Table 4, we investigate the performance of our bimodal AcFormer by individually removing one modality. Our findings indicate a significant performance drop when text modality is removed, suggesting its predominant role in the MER task. This is due to the fact

Configs	CMU-MOSI		CMU-MOSEI		IEMOCAP	
	MAE(↓)	Corr(↑)	MAE(↓)	Corr(↑)	Acc.(↑)	F1(↑)
Role of Each Modality (AcFormer♠ (B))						
Video only	1.651	0.119	1.016	0.103	68.53	66.24
Audio only	1.231	0.296	0.944	0.262	70.43	68.45
Text only	0.989	0.613	0.673	0.665	82.34	79.67
V+A	1.237	0.313	0.915	0.376	74.22	72.53
V+T	0.784	0.705	0.600	0.714	85.63	83.26
A+T	0.766	0.746	0.568	0.759	86.21	84.12
V+A+T	**0.715**	**0.794**	**0.531**	**0.786**	**87.03**	**85.55**
Effect of Pre-fusion Alignment Modules (AcFormer♣ (B))						
+CMA	0.726	0.783	0.535	0.778	85.67	83.60
+IMC	0.782	0.757	0.562	0.713	83.37	82.16
+(CMA,IMC)	0.711	0.806	0.513	0.793	86.45	85.21
+(CMA,IMC,VAT-M)	**0.703**	**0.815**	**0.518**	**0.802**	**86.98**	**85.53**
Impact of Aggregation Operators (AcFormer♠ (B))						
CONCAT	0.716	0.790	**0.528**	0.783	87.01	85.40
AVG	0.718	0.787	0.532	0.778	**87.12**	**85.61**
MAX	0.724	0.763	0.537	0.772	86.97	85.18
SUM	**0.715**	**0.794**	0.531	**0.786**	87.03	85.55

Table 4: Ablation studies for pre-fusion alignment modules and candidate aggregators. T=text, A=audio and V=video.

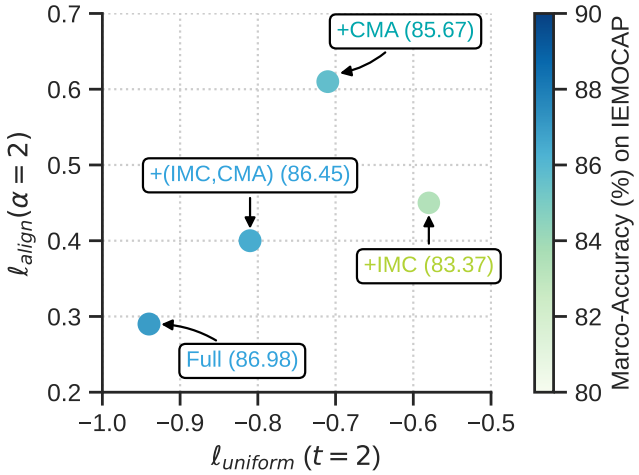

Figure 2: $\ell_{align}-\ell_{uniform}$ plot of models. Color of points and numbers in brackets represent average MER performance.

that visual and audio features are more prone to spatio-temporal redundancy and noise, compared to textual features. Furthermore, we observe that the trimodal AcFormer achieves the best results, confirming that each modality plays a distinct role in our model. We also study the impact of the proposed modality alignment modules in Table 4. First, we consider AcFormer with removing all alignment modules as the baseline. Then, we add the CMA module and observe a significant performance improvement in AcFormer. This suggests that explicit alignment between modalities is instrumental in learning stronger multimodal representations. Similar gains are found upon adding the IMC module, as it ensures that similar inputs from the same modality remain close in the embedding space, thereby greatly reducing the impact of data noise in each modality. The most significant improvement is achieved when both modules are

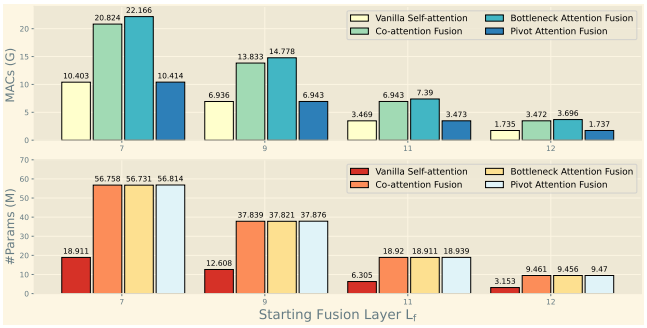

Figure 3: The computational overhead of different modality fusion strategies varies with different fusion layers. Here we unify the number of pivot/bottleneck tokens to 12, the attention heads to 8, and the embedding dimension to 768.

added, revealing the synergetic effect of CMA and IMC in reducing modality gap and redundancy. The insertion of VAT-M also makes marginal contributions to the model performance. Finally, we see that AcFormer is not sensitive to changes in aggregator functions.

4.5.2 ***Quality of Modality Representation.*** Following [50], we also plot the *uniformity* and *alignment* of our AcFormer variants in Fig. 2 to assess the quality of representations. In general, we find that models with better alignment and uniformity achieve better performance, which is in line with [50]. We can also see that 1) IMC works well in improving alignment, while CMA encourages uniform distribution of diverse unimodal embeddings; and 2) VAT-M further enhances both the alignment and uniformity. Fig. 4 illustrates the t-SNE plots [48] of modality representations. It is seen that, before applying alignment modules, modality representations appear scattered and poorly organized. After adding IMC, meaningful clusters can be formed within each modality, but the features between modalities are still hard to separate. Further adding CMA leads to separation of features from different modalities belonging to the same class, which means that the learned representations can capture more discriminative and semantically consistent information.

4.5.3 ***Fusion Strategy.*** As illustrated in Fig. 5A, we investigate how the model performance varies with the number of pivot tokens, *i.e.* P. We conduct experiments with P = [4, 8, 12, 24, 48, 96, 128]. The MER performance keeps increasing as P increases from 4 to 12, but does not continue to improve as **P** further increases. Therefore, to strike a balance between performance and efficiency, we set **P** = 12 as the default number of pivot tokens. Next, we will unveil the effect of our modality fusion strategy by the following comparative analysis of several fusion strategies:

- **Vanilla Self-Attention** [49]: This strategy applies vanilla self-attention between all concatenated latent units per layer.
- **Co-Attention** [28]: The fusion latent units are updated via pairwise attention with latent units from all other modalities.
- **Bottleneck Attention Fusion** [33]: All cross-modal attention is required to pass through bottleneck fusion tokens in a **coupled** modal interaction manner.
- **Pivot Attention Fusion (Ours)**: Similar to Bottleneck Fusion, but in a **decoupled** modal interaction fashion.

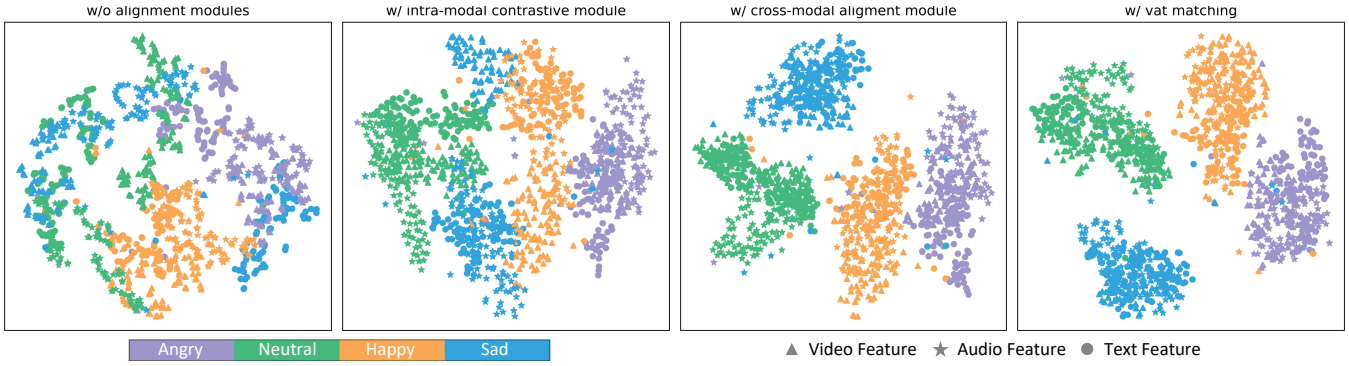

Figure 4: t-SNE plot of modality representation before and after applying the alignment modules on the MOSEI dataset. The insertion of IMC, CMA and VAT-M consistently leads to better separation of dissimilar samples and clustering of similar samples.

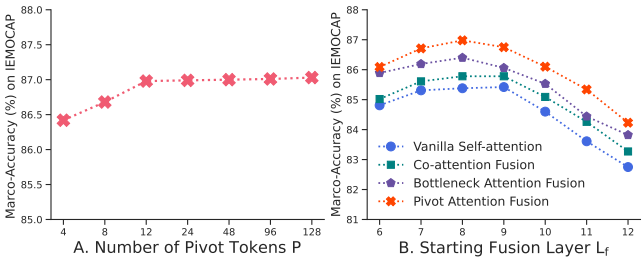

Figure 5: Ablation studies on the number of pivot tokens, different fusion layers and mulitmodal fusion strategies.

Note that these fusion strategies only describe attention flow between latent units within a layer. To investigate the impact of restricting cross-modal attention to layers after a fixed fusion layer L_f, we show the performance of AcFormer variants varying with different fusion layers and different fusion strategies in Fig. 5B. We can observe that: i) our pivot attention (PA) fusion always performs better than the self-attention (SA), co-attention (CA) and bottleneck attention (BA) fusion at the same starting fusion layer. This reveals the effectiveness of making the pivot token a bridge for cross-modal information exchange; ii) when $L_f = 8$, AcFormer achieves optimal performance, suggesting that it allows earlier layers to focus on learning unimodal features while still benefiting from restricting cross-modal connections to higher layers.

Complexity Analysis. Sticking to the previous notations, let n denote the sequence length of the concatenated tokens and d denote the feature dimension. Assume $n = \mathrm{M} + \mathrm{N} + \mathrm{L}$. Denote by p the number of pivot tokens. The per-layer time complexities of SA, CA, BA and PA are $O(n^2d+nd^2)$, $O((\mathrm{M}^2+\mathrm{N}^2+\mathrm{L}^2+n^2)d+nd^2)$, $O((\mathrm{M}^2+\mathrm{N}^2+\mathrm{L}^2+pn)d+nd^2)$ and $O(pnd+pd^2)$, respectively. Detailed derivations can be found in Sup. G. Since p is much smaller than n, our model enjoys relatively low computational overhead. We test different fusion methods using a standard multimodal input by setting the batch size to 1, $d=768$, $\mathrm{M}=196$ (8 sampled video frames), $\mathrm{N}=98$ (time frames of 1s audio), $\mathrm{L}=256$ and $p=12$. The number of MACs and parameters of each fusion method are presented in Fig. 3. We find that, despite having comparable parameter counts between PA, CA, and BA when operating on the same fusion layers, PA requires fewer GMACs. This strongly demonstrates the computational efficiency of our multimodal feature fusion mechanism.

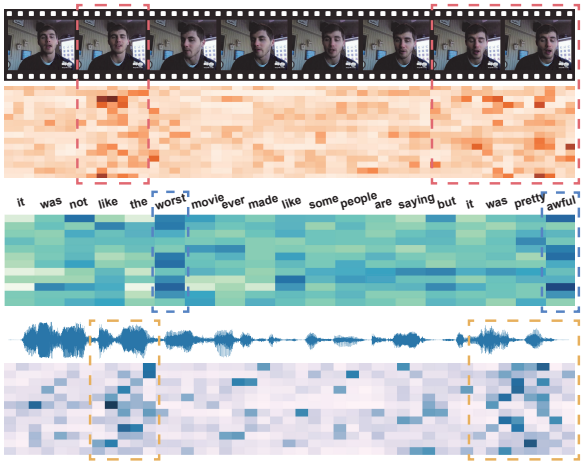

Figure 6: Attention maps computed between the pivot tokens and the latent units of each modality (derived from Eq. (6)).

4.5.4 ***Visualization of Pivot Attention.*** Fig. 6 depicts the attention maps in the last pivot attention block, computed by viewing pivot tokens as *queries* and latent tokens of each modality as *keys*. We can observe that our pivot attention is capable of successfully associating with emotion-related video clips, textual words, and speech segments. Moreover, from the distribution of attention weights, it can be inferred that these emotion-related multimodal sequence segments are not aligned at the word-level, often accompanied by temporal delays, where a single word may correspond to several video frames. Our pivot attention is capable of capturing such long- and short-range dependencies across modalities, demonstrating its effectiveness in multimodal feature fusion.

5 CONCLUSION

In this paper, we present AcFormer, an aligned and compact Transformer for multimodal sentiment analysis. The core ingredient of AcFormer lies in i) *pre-fusion alignment* modules that aligns different modality streams using both intra-modality and inter-modality contrastive learning, and ii) *pivot attention* for multimodal feature fusion by modulating cross-attention flow to a small set of pivot tokens. Experimental results show that AcFormer outperforms previous SOTA methods with little computational overhead.

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS* 33 (2020), 12449–12460.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *WACV*. IEEE, 1–10.

[4] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* 226 (2021), 107134.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42 (2008), 335–359.

[6] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815* (2019).

[7] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *EMNLP-IJCNLP*. 5647–5657.

[8] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 163–171.

[9] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379* (2018).

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR, 1597–1607.

[11] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[14] Thomas Drugman and Abeer Alwan. 2019. Joint robust voicing detection and pitch estimation based on residual harmonics. *arXiv preprint arXiv:2001.00459* (2019).

[15] Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. 2011. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 3 (2011), 994–1006.

[16] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[17] Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically Adjust Word Representations Using Unaligned Multimodal Information. In *ACM Multimedia*. 3394–3402.

[18] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618* (2019).

[19] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM Multimedia*. 1122–1131.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*. 9729–9738.

[21] Penn Phonetics Laboratory. 2013. p2fa-vislab. https://github.com/ucbvislab/p2fa-vislab/. A script for audio/transcript alignment.

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 12888–12900.

[23] 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. *NeurIPS* 34 (2021), 9694–9705.

[24] Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In *EMNLP*. 150–161.

[25] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*. 3042–3051.

[26] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *ACL*. 2247–2256.

[27] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS* 32 (2019).

[29] 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 *CVPR*. 2554–2562.

[30] Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *ACL*. 481–492.

[31] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *AAAI*, Vol. 34. 164–172.

[32] Sijie Mai, Songlong Xing, and Haifeng Hu. 2019. Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia* 22, 1 (2019), 122–137.

[33] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *NeurIPS* 34 (2021), 14200–14213.

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

[35] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proc. Interspeech 2019* (2019), 2613–2617.

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[37] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, Vol. 33. 6892–6899.

[38] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*. 2539–2544.

[39] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

[40] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 338–353.

[41] Erika L Rosenberg and Paul Ekman. 2020. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

[42] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In *ACM Multimedia*. 3722–3729.

[43] 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* (2023).

[44] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *AAAI*, Vol. 34. 8992–8999.

[45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*. PMLR, 10347–10357.

[46] 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 *ACL*, Vol. 2019. NIH Public Access, 6558.

[47] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018).

[48] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* 30 (2017).

[50] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*. PMLR, 9929–9939.

[51] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, Vol. 33. 7216–7223.

[52] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges.

Artificial Intelligence Review 55, 7 (2022), 5731–5780.

[53] Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics*. 4730–4738.

[54] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled Representation Learning for Multimodal Emotion Recognition. In *ACM Multimedia*. 1642–1651.

[55] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP*. 1103–1114.

[56] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *AAAI*, Vol. 32.

[57] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *AAAI*, Vol. 32.

[58] 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 (2016), 82–88.

[59] 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 *ACL*. 2236–2246.