MAF-Net: Modality-Aligned Fusion with Modular Attention for Multimodal Sentiment Analysis

1st Yanbo Wang
College of Intelligent Science and Technology
Inner Mongolia University of Technology
Hohhot, China
20241100138@imut.edu.cn

2nd Qing-Dao-Er-Ji Ren*
College of Intelligent Science and Technology
Inner Mongolia University of Technology
Hohhot, China
renqingln@imut.edu.cn

***Abstract*—Multimodal Sentiment Analysis (MSA) aims to decipher human emotional states by synthesizing linguistic, acoustic, and visual signals. Despite recent advancements, effective cross-modal interaction remains hindered by two fundamental challenges: the inherent semantic gap between dense textual representations and redundant, noisy audio-visual signals, and the phenomenon of unimodal dominance, where models over-rely on the dominant modality (text) at the expense of subtle non-verbal cues. To address these issues, we propose MAF-Net (Modality-Aligned Fusion with Modular Attention). First, we introduce a Geometric Semantic Alignment (GSA) mechanism. Unlike traditional linear mappings, GSA leverages Gram matrix-based basis vectors to project non-verbal features into a unified textual manifold, effectively filtering low-level noise while preserving essential geometric structures. Second, to counter contribution imbalance, we design Modular Masked Interactive Fusion (MMIF). This strategy employs a decoupled attention masking scheme to independently regulate self-modal and cross-modal information flows, preventing the suppression of recessive modalities during fusion. Extensive experiments on standard benchmarks (CMU-MOSI, CMU-MOSEI) and a newly constructed Mongolian MSA dataset (MG-SIMS) demonstrate that MAF-Net achieves state-of-the-art performance. Our results highlight the framework's robustness in complex interaction scenarios and its superior transferability to low-resource language settings.**

***Index Terms*—Multimodal Sentiment Analysis, Geometric Semantic Alignment, Cross-modal Fusion, Unimodal Dominance, Low-resource Languages**

I. Introduction

With the proliferation of social media and online video platforms, multimodal data—encompassing text, audio, and video—has emerged as the dominant medium for information dissemination on the internet. Multimodal Sentiment Analysis (MSA) aims to identify and comprehend human emotional states by integrating linguistic, acoustic, and visual information, holding significant value in applications such as human-computer interaction, public opinion monitoring, and intelligent customer service. While mainstream models have achieved remarkable strides on general language datasets, efficiently fusing heterogeneous modal features and resolving cross-modal semantic discrepancies remains a critical challenge, particularly for specific languages like Mongolian.

Despite the momentum provided by deep learning in MSA research, constructing efficient cross-modal interaction mechanisms continues to face severe difficulties. These challenges primarily center on the inefficiency of interaction caused by the cross-modal semantic gap and information redundancy, as well as the contribution imbalance during the modal fusion process.

First, a distinct semantic gap and disparity in information density exist between different modalities. The textual modality typically encapsulates highly abstract and dense semantic information. In contrast, acoustic and visual modalities consist of continuous spatiotemporal signals; while they contain rich non-verbal emotional cues (e.g., intonation, micro-expressions), they are also accompanied by substantial low-level redundancy and background noise. Existing cross-modal interaction methods, such as standard cross-attention mechanisms, often struggle to precisely capture deep correlations with textual semantics while simultaneously suppressing non-verbal noise. This "coarse-grained" direct interaction frequently leads to a dilemma: either critical emotional details in non-verbal modalities are submerged by high-density textual semantics, or irrelevant noise from audio-visual signals interferes with textual feature expression. Consequently, this limits the model's discriminative capability for complex emotional states, especially subtle sentiments like neutrality or sarcasm.

Second, in the feature fusion stage, existing approaches predominantly rely on simple concatenation or similarity-based weighted fusion. These methods often fail to effectively differentiate the contributions of "dominant" versus "recessive" modalities, causing unique emotional details in weaker modalities to be overshadowed and hindering the realization of complementary advantages across modalities.

To address these challenges, we propose a novel multimodal sentiment analysis framework. By integrating fine-grained alignment with sophisticated fusion strategies, our method significantly enhances model robustness in complex interaction scenarios. Furthermore, we extend our framework to the Mongolian MSA dataset, achieving state-of-the-art performance. These results demonstrate the strong generalizability and transferability of our model, particularly in low-resource language settings.

To resolve modal semantic heterogeneity and bias, we introduce a Geometric Semantic Alignment (GSA) mechanism. Unlike traditional linear mapping, we innovatively utilize the Gram matrix of textual token representations to extract Basis Vectors from the semantic space. These vectors serve as a reference benchmark to map acoustic and visual features into a unified textual semantic space. This approach significantly

*Corresponding author: renqingln@imut.edu.cn

reduces low-level redundancy in non-textual modalities while preserving their original information.

Building upon feature alignment, we further propose a Modular Masked Interactive Fusion (MMIF) strategy to address the contribution imbalance between strong and weak modalities. While existing fusion methods risk "averaging" features, MMIF employs modular attention masks to decompose the attention flow into independent self-modal and cross-modal streams. By regulating information interaction through a learnable masking strategy, MMIF prevents dominant modalities from suppressing weaker ones. This ensures that the core semantic structure of the text is maintained while fully exploiting the complementary value of non-verbal modalities, guaranteeing the balanced utilization of tri-modal information in emotional reasoning.

The main contributions of this paper are summarized as follows:

We propose the Geometric Semantic Alignment (GSA) mechanism, which achieves efficient cross-modal semantic alignment via Gram matrix-based basis vector mapping, effectively bridging the semantic gap between heterogeneous modalities.

We design the Modular Masked Interactive Fusion (MMIF) strategy, which effectively alleviates the information imbalance between strong and weak modalities during fusion through a decoupled attention masking strategy.

We construct a specialized Multimodal Sentiment Analysis model for Mongolian. Experimental results demonstrate that our method not only improves sentiment recognition accuracy but also significantly enhances model robustness and the ability to capture subtle emotional features.

II. Related Work

A. Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) aims to interpret human emotional states by synthesizing heterogeneous signals from text, audio, and visual modalities. Early approaches predominantly relied on Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to model temporal sequences [1], [2]. For instance, the Memory Fusion Network (MFN) [3] utilized a multi-view gated memory to align interactions over time. However, these recurrent architectures often struggle with capturing long-range dependencies due to the vanishing gradient problem, limiting their ability to model complex emotional dynamics over extended contexts.

In recent years, the Transformer architecture [4] has become the de facto standard in MER due to its superior ability to capture global dependencies via self-attention mechanisms. Pioneering works like the Multimodal Transformer (MulT) [5] introduced cross-modal attention to implicitly align unaligned sequences, achieving significant performance gains. Subsequently, methods like MISA [6] focused on learning invariant and specific representations by projecting modalities into distinct subspaces.

Despite these advancements, a critical gap remains in handling modal heterogeneity. Most existing Transformer-based methods employ coarse-grained interaction strategies, such as direct feature concatenation or global cross-attention, which implicitly assume that all input segments are equally informative [5], [7]. However, non-verbal modalities (audio and vision) are inherently unstructured and contain substantial redundancy and background noise (e.g., unvoiced segments, irrelevant visual background) [8]. Directly fusing these noisy features without effective filtering or selection often dilutes the core semantic information, making it difficult for the model to extract discriminative emotional cues. This limitation underscores the need for more refined mechanisms to filter noise and align heterogeneous features effectively.

B. Imbalanced Multimodal Learning

Recent studies have highlighted the issue of imbalanced multimodal learning, where models tend to favor certain modalities (e.g., text) over others, negatively impacting overall performance [9], [10]. Various strategies have been proposed to address this challenge [11]–[13]. Peng et al. [10]introduced a gradient modulation strategy that dynamically adjusts the contributions of different modalities during training.

However, most existing approaches address this imbalance primarily at the optimization level (e.g., gradient adjustment) or the sample level. These methods often fail to explicitly model the fine-grained interactions within the attention mechanism itself, potentially leaving the internal feature representations dominated by the strong modality. In contrast, we propose a structural solution via the Modular Masked Interactive Fusion (MMIF) strategy. Instead of merely adjusting gradients, we design decoupled attention masks to explicitly regulate the attention flow. This ensures that the "weak" modalities (audio and video) are not suppressed but are actively integrated into the semantic space through independent self-modal and cross-modal attention streams, thereby achieving intrinsic feature-level balance.

C. Multimodal Feature Fusion Strategies

A major direction involves improving multimodal integration to leverage complementary information [14]. Traditional early fusion relied on concatenation of features [15], while Zadeh et al. [16] later introduced tensor fusion to capture high-order interactions via trilinear decomposition. More recently, transformer-based methods like MulT [5] have utilized cross-modal attention to model correlations between modalities. However, these models typically apply a global fusion mechanism, neglecting instance-level variations in modality importance. In a separate work stream, Hazarika et al. [6] proposed disentangle features into modality-invariant and modality-specific components to learn refined representations. Consequently, modality decoupling has become a key strategy for handling modality heterogeneity.

III. Methods

A. Model Overview

Figure 1 illustrates the overall architecture of the proposed Modality-Aligned Fusion with Modular Attention (MAF)

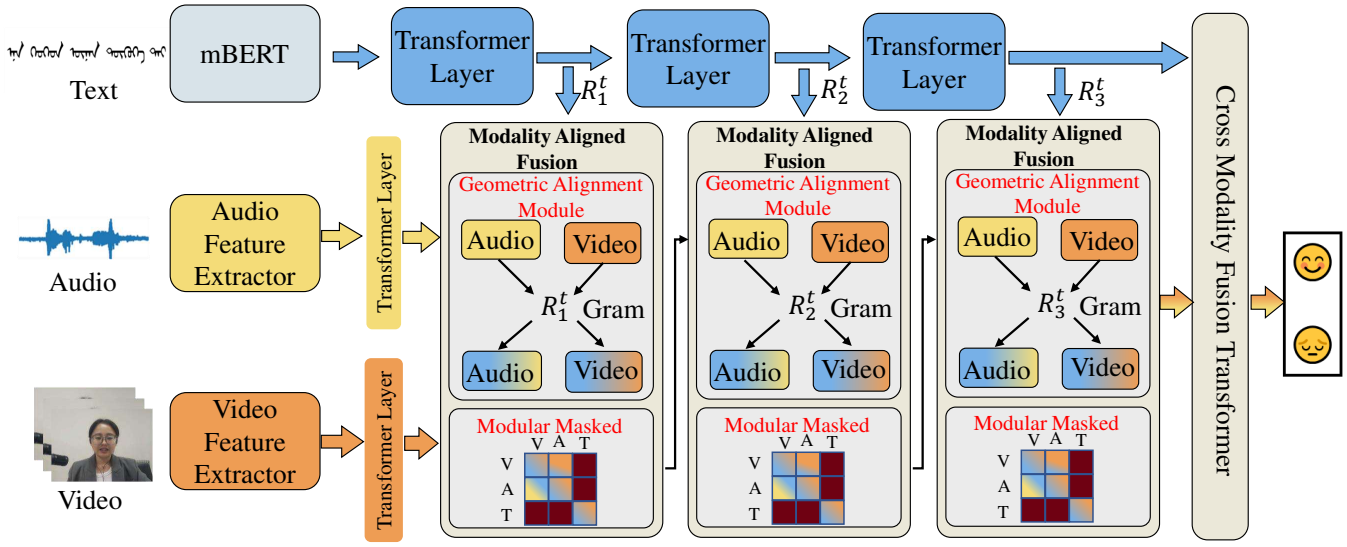

Fig. 1. The overall architecture of the proposed MAF for multimodal sentiment analysis(MSA). With the multimodal input, we first apply three Transformer layers to embed modality features with low redundancy. Then, we employ a three-layer cascaded MAF module to progressively align and fuse the extracted primary features of each modality under the guidance of the text modality. Wherein, each MAF module is composed of a GSA module and an MMIF module. Finally, a Cross-modality Fusion Transformer is utilized to fuse the textual modality with the joint audiovisual features, thus obtaining a complementary and joint representation for MSA.

framework. As shown, MAF-Net first extracts unimodal features from the raw inputs. Then, these features are processed by a cascade of three MAF Blocks to perform a deep interaction. Within each block, a Geometric Semantic Alignment module is first employed to align heterogeneous features into a unified semantic space, followed by a Modular Masked Fusion module to dynamically integrate them. Finally, the output of the last MAF Block is used to generate the final prediction for MSA.

B. Unimodal Embedding

Consistent with prior literatures [8], [17], we extract features using mBERT, Librosa, and OpenFace. We formulate the input for modality $m \in \{t, v, a\}$ as a sequence $I_m \in \mathbb{R}^{T_m \times d_m}$, where T_m denotes the sequence length and d_m represents the feature dimension.While these dimensions vary by dataset, for the MOSI benchmark, the sequence lengths are unified at $T_m = 50$, with feature dimensions of $d_t = 768$, $d_v = 20$, and $d_a = 5$.

We utilize a two-phase framework for unimodal embedding.To project heterogeneous modality features into a unified subspace, we employ modality-specific Transformer encoders. Specifically, we define a set of learnable latent tokens $R_0^m \in \mathbb{R}^{T \times d_m}$ for each modality. These tokens are concatenated with the input I_m and processed to distill essential semantic information:

$$R_1^m = E_0^m \left(\text{concat}\left(R_0^m, I_m\right); \theta_{E_0^m}\right) \in \mathbb{R}^{T \times d} \tag{1}$$

where R_1^m denotes the unified feature representation, and E_0^m represents the feature extractor parameterized by $\theta_{E_0^m}$. We adopt a single-layer Vision Transformer (ViT) [18]architecture for E_0^m. In our implementation, we set the token length $T = 8$ and the unified dimension $d = 128$. By constraining the information flow to these low-dimensional tokens, this design acts as an information bottleneck, effectively filtering out sentiment-irrelevant noise and redundancy while maintaining computational efficiency.

C. Geometric Semantic Alignment

After unimodal feature extraction, we propose a Geometric-Semantic Alignment (GSA) module. The core objective of this module is to address the semantic misalignment problem between heterogeneous modalities before fusion, ensuring that non-verbal features (audio and video) are effectively mapped into the more robust textual semantic space.

a) Gram-based Feature Projection

To mitigate the inherent semantic heterogeneity across video, audio, and text modalities, we propose a projection mechanism that maps non-verbal tokens into the textual manifold using basis vectors derived from the Gram matrix.

Existing approaches often suffer from significant feature distribution shifts between visual/acoustic signals and textual information. Consequently, direct attention mechanisms frequently fail to capture authentic cross-modal correspondences. To address this limitation, we exploit the Gram matrix of textual tokens to construct a spatial basis. This basis serves as a semantic transformation medium, effectively compressing textual semantics to guide the alignment of other modalities.

As illustrated in Figure 2, the proposed alignment framework comprises two specialized components: the V-Aligner and A-Aligner. These modules are dedicated to projecting video and audio representations, respectively, into the unified textual coordinate system.

b) Formulation

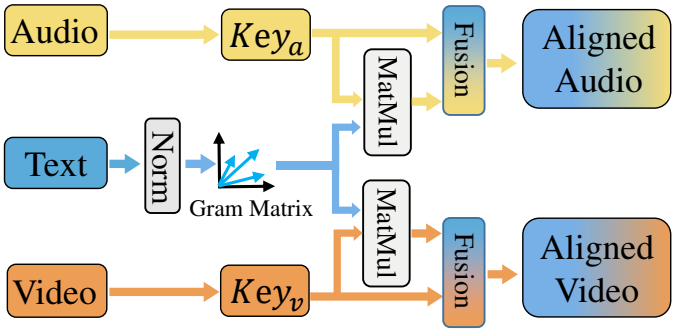

Fig. 2. An illustration of the Geometric Semantic Alignment (GSA) module. Guided by the text modality, the module aligns the audio and video features into a unified semantic space by calculating the Gram Matrix of the textual features.

Formally, for the textual modality ($m = t$), we construct a spatial basis derived from the normalized Gram matrix $\|G^m\| \in \mathrm{R}^{d\times d}$. Here, each element G^{m}_{ij} represents the inner product correlation, calculated as:

$$G^m_{ij} = \sum_{k=1}^{N_m} K^m_{ik} K^m_{kj} = (K^m)^\top K^m \tag{2}$$

where $\mathbf{K}_m \in \mathbb{R}^{N_m \times d}$ denotes the Key states of the textual tokens, and N_m represents the token sequence length. By incorporating the spatial basis defined by the Gram matrix, the model effectively encapsulates the intra-modal geometric relationships within the text, establishing a unified semantic reference frame for the video and audio modalities.

c) Alignment

Acting as a cross-modal transformation operator, the normalized textual Gram matrix is employed to efficiently project tokens from the video (v) and audio (a) modalities into the textual manifold. We define this core mapping function as $f : \mathbb{R}^d \to \mathbb{R}^d$. The aligned representations are computed as follows:

$$K^{\bar{v}\to t} = K^{\bar{v}} \left\|G^{\mathrm{t}}\right\| \tag{3}$$

$$K^{\bar{\mathrm{a}}\to t} = K^{\bar{\mathrm{a}}} \left\|G^{t}\right\| \tag{4}$$

where $K^{\bar{\mathrm{v}}}$ and $K^{\bar{\mathrm{a}}}$ represent the Key states of the visual and acoustic modalities, respectively. These projected tokens are subsequently fused with the original textual tokens to enhance cross-modal semantic consistency and achieve robust feature alignment. Notably, this alignment operation maintains linear complexity with respect to the token sequence length, ensuring high computational efficiency.

D. Modular Masked Interactive Fusion

Attention masks control the information flow across modalities and encode structural priors for multimodal transformers. To address the natural synchronization between modalities and emotion-specific cross-modal dependencies, we design the **Modular Masked Interactive Fusion (MMIF)** . This mechanism is instantiated as a composite attention mask $\mathcal{M}$ consisting of three components: intra-modality positional decay $\mathcal{M}_{\text{pos}}$, audio-visual synergy enhancement $\mathcal{M}_{\text{sync}}$, and emotion alignment enhancement $\mathcal{M}_{\text{emo}}$.

The attention mechanism incorporating our MMIF is formulated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \mathcal{M}\right) V \tag{5}$$

where $\mathcal{M} = \mathcal{M}_{\text{pos}} + \mathcal{M}_{\text{sync}} + \mathcal{M}_{\text{emo}}$.

Intra-Modality Positional Decay. To model temporal dependencies within each modality sequence, we apply a distance-aware positional bias with a decay rate β. For tokens within the same modality, the attention score to distant positions is penalized proportionally to their absolute distance:

$$\mathcal{M}^{ij}_{\text{pos}} = \begin{cases} p_{\text{base}} - \beta \cdot |i-j| & \text{if } m_i = m_j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where m_i denotes the modality of token i, and $p_{\text{base}} = 0$ is the baseline bias. Note that $\mathcal{M}_{\text{pos}}$ is modality-specific (indicated by superscripts $\mathcal{M}^{\mathrm{V}}_{\text{pos}}$, $\mathcal{M}^{\mathrm{A}}_{\text{pos}}$, $\mathcal{M}^{\mathrm{T}}_{\text{pos}}$) because video, audio, and text each possess varying sequence lengths and temporal characteristics, necessitating independent positional encoding strategies. This mechanism prevents attention collapse and establishes local temporal coherence within each distinct modality stream.

Audio-Visual Synergy Enhancement. Motivated by the natural synchronization between visual and auditory signals in emotional expression (e.g., facial expressions aligning with vocal tone, or body language with prosody), we explicitly enhance the bidirectional cross-modal attention between video and audio tokens through a positive bias:

$$\mathcal{M}^{ij}_{\text{sync}} = \begin{cases} \gamma & \text{if } (m_i, m_j) \in \{(\mathrm{V},\mathrm{A}), (\mathrm{A},\mathrm{V})\} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\gamma > 0$ is a synergy boost factor. This positive bias amplifies the attention weights following the softmax normalization, thereby facilitating a more fine-grained audio-visual fusion. This design exploits the complementary nature of visual and acoustic emotional cues, such as the concurrent manifestation of tense facial expressions and aggressive vocal patterns in anger. Unlike $\mathcal{M}_{\text{pos}}$, $\mathcal{M}_{\text{sync}}$ applies uniformly to all video-audio interactions with a shared parameter γ, reflecting the universal principle of audiovisual synchronization.

Emotion Alignment Enhancement. In emotion analysis, textual semantics often exhibit inconsistencies with non-verbal cues, particularly in complex expressive scenarios involving sarcasm, emotional suppression, or deception. To enable the model to effectively detect such cross-modal emotional conflicts, we **unidirectionally** enhance text-to-perceptual attention using a positive bias:

$$\mathcal{M}^{ij}_{\text{emo}} = \begin{cases} \delta & \text{if } (m_i, m_j) \in \{(\mathrm{T},\mathrm{V}), (\mathrm{T},\mathrm{A})\} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $\delta > 0$ is an emotion alignment factor. Critically, this enhancement is **asymmetric**: it is applied solely when

text tokens query perceptual modalities (T→V, T→A), and not in the reverse direction (V→T, A→T). This asymmetric architecture simulates the cognitive mechanism underlying human emotion comprehension: when linguistic expressions prove ambiguous, we actively seek verification and context from perceptual cues, such as facial expressions and vocal intonation.

Complete Attention Structure. The complete attention matrix integrating our MMIF exhibits a hierarchical structural pattern. This guarantees temporal coherence within each individual modality while achieving targeted cross-modal interactions optimized for emotion understanding:

$$A_{\text{MMIF}} = \begin{pmatrix} A^{\text{VV}} + \mathcal{M}^{\text{V}}_{\text{pos}} & A^{\text{VA}} + \mathcal{M}_{\text{sync}} & A^{\text{VT}} \\ A^{\text{AV}} + \mathcal{M}_{\text{sync}} & A^{\text{AA}} + \mathcal{M}^{\text{A}}_{\text{pos}} & A^{\text{AT}} \\ A^{\text{TV}} + \mathcal{M}_{\text{emo}} & A^{\text{TA}} + \mathcal{M}_{\text{emo}} & A^{\text{TT}} + \mathcal{M}^{\text{T}}_{\text{pos}} \end{pmatrix} \tag{9}$$

where A^{mn} denotes the raw attention scores from modality m to modality n. This matrix structure embodies three core design principles: **(1) Diagonal blocks** contain modality-specific positional decay ($\mathcal{M}^{\text{V/A/T}}_{\text{pos}}$), providing independent positional modeling capabilities for the temporal sequences of video, audio, and text; **(2) Video-audio blocks** incorporate bidirectional synergy enhancement ($\mathcal{M}_{\text{sync}}$), leveraging the natural synchronization of audiovisual signals to promote the complementary fusion of perceptual features; and **(3) Text-to-perceptual blocks** apply unidirectional emotion alignment enhancement ($\mathcal{M}_{\text{emo}}$). This allows the semantic layer to actively integrate perceptual evidence to resolve emotional ambiguity, while the reverse directions maintain standard attention mechanisms to prevent potential modal over-coupling caused by bidirectional reinforcement.

IV. Experiment

A. Datasets

We conduct extensive experiments on three standard multimodal sentiment analysis benchmarks: MOSI [19], MOSEI [20], as well as MG-SIMS, our self-constructed Mongolian multimodal sentiment analysis dataset.

- **CMU-MOSI.**The CMU-MOSI dataset is a popular benchmark dataset in MSA research. This dataset is a collection of YouTube monologues. MOSI contains 2199 subjective words-video clips. These utterances are artificially labeled as consecutive opinion scores between -3 to 3, where -3/+3 represents strong negative/positive sentiment.
- **CMU-MOSEI.**The CMU-MOSEI dataset is an improvement on MOSI. It contains 23,454 YouTube monologues video segments covering 250 distinct topics from 1,000 distinct speakers. Each utterance also has sentiment consecutive opinion scores between -3 to 3.
- **MG-SIMS.**The MG-SIMS dataset consists of 2,100 Mongolian multimodal samples. It includes seven discrete emotions: happiness, anger, sadness, surprise, fear, disgust, and neutrality (representing emotionally steady speech). Each emotion category contains 300 sentences.

B. Evaluation Metric

Following the previous work [8], [22], we present our experimental findings in two distinct formats: classification and regression. In terms of classification, we used several evaluation metrics, i.e., Weighted F1 score(F1-Score), binary classification accuracy (Acc-2), three classification accuracy(Acc-3), five classification accuracy (Acc-5), seven classification accuracy (Acc-7), mean absolute error (MAE), and the correlation of the model's prediction with human (Corr). Specifically, for the MOSI, MOSEI and MG-SIMS, we compute Acc-2 and F1-score in two configurations: negative / non-negative (including zero) and negative / positive (excluding zero). Moreover, we include additional metrics such as Acc-5 and Acc-7. For MG-SIMS, we use Acc-7, MAE, F1 and Corr to evaluate. Regarding regression, we report MAE and Corr. In all metrics except MAE, higher values indicate better performance.

C. Baselines

In order to verify the superiority of our proposed MAF, we conduct an experimental comparison with the following state-of-the-art baseline models, and the relevant results have been presented in Table I and Table II:

- **Utterance-vector fusion approaches that use tensor-based fusion and low-rank variants:** TFN [16], LMF [22].
- **Models which Learn invariant and specific representations through feature decomposition:** MISA [6] , FDMER [23], MFSA [24], ConFEDE [25].
- **Models which utilize attention and transformer modules to improve token representations using non-verbal signals:** MulT [5], PMR [26], ALMT [17], DEVA [21].

D. Performance Comparison

Table Iand Table II list the comparison results of our proposed method and state-of-the-art methods on the MOSI, MOSEI, and the newly constructed Mongolian dataset MG-SIMS, respectively. As shown in Table I, the proposed MAF-Net obtained state-of-the-art performance across the majority of metrics on standard benchmarks.

On the task of more difficult and fine-grained sentiment classification (Acc-5 and Acc-7), our model achieves remarkable improvements. For example, on the MOSI dataset, MAF-Net achieved an absolute improvement of 3.32% on Acc-5 (55.10% vs. 51.78%) and 2.95% on Acc-7 (49.27% vs. 46.32%) compared to the strong baseline DEVA. It demonstrates that by explicitly aligning geometric structures via the Geometric Semantic Alignment (GSA) mechanism, our model can effectively filter low-level noise while preserving subtle non-verbal cues, which are critical for distinguishing complex emotional states.

Furthermore, on the large-scale MOSEI dataset, MAF-Net consistently outperforms baselines, achieving the lowest MAE (0.518) and the highest correlation (0.786). This confirms that the Modular Masked Interactive Fusion (MMIF) strategy

TABLE I
COMPARISON ON MOSI AND MOSEI DATASETS. * REPRESENTS RESULTS OBTAINED FROM DEVA [21]. MODELS WITH † ARE REPRODUCED UNDER THE SAME CONDITIONS. BEST RESULTS ARE MARKED IN BOLD.

Methods	MOSI						MOSEI					
	Acc-2	Acc-5	Acc-7	F1	MAE	Corr	Acc-2	Acc-5	Acc-7	F1	MAE	Corr
TFN*	77.99/79.08	-	34.46	77.95/79.11	0.947	0.673	78.50/81.89	-	51.60	78.96/81.74	0.572	0.714
LMF*	77.90/79.18	-	33.82	77.80/79.15	0.950	0.651	80.54/83.48	-	51.59	80.94/83.36	0.575	0.716
MulT*	79.71/80.98	42.68	36.91	79.63/80.95	0.879	0.702	81.15/84.63	54.18	52.84	81.56/84.52	0.559	0.733
ICCN	-/83.07	-	39.01	-/83.02	0.862	0.714	-/84.18	-	51.58	-/84.15	0.565	0.713
MISA*	81.84/83.54	47.08	41.37	81.82/83.58	0.776	0.778	80.67/84.67	53.63	52.05	81.12/84.66	0.557	0.751
MAG-BERT	82.37/84.43	-	43.62	82.50/84.61	0.727	0.781	82.51/84.82	-	52.67	82.77/84.71	0.543	0.755
PMR	-/82.40	-	40.60	-/82.10	-	-	-/83.10	-	51.80	-/82.80	-	-
MFSA	-/83.3	-	41.1	-/83.7	0.856	0.722	-/83.8	-	53.2	-/83.6	0.574	0.724
FDMER	-/84.6	-	44.1	-/84.7	0.724	0.788	-/86.1	-	54.1	-/85.8	0.536	0.773
Self-MM†	82.54/84.45	52.22	45.56	82.46/84.44	0.719	**0.794**	82.09/84.76	53.54	53.65	82.43/84.67	0.535	0.761
ALMT†	83.24/85.37	50.29	44.75	83.41/85.46	0.738	0.776	82.34/85.94	55.05	53.32	81.85/85.93	0.534	0.771
ConFEDE	84.17/85.52	-	42.27	84.13/85.52	0.742	0.784	81.65/85.82	-	54.86	82.17/85.83	0.522	0.780
DEVA†	83.62/85.73	51.06	45.62	84.03/85.60	0.730	0.774	81.83/85.41	55.60	53.14	81.89/85.97	0.556	0.763
MAF-Net	**84.40/86.13**	**55.10**	**49.27**	**84.25/86.03**	**0.702**	0.791	**83.65/86.58**	**55.84**	**55.12**	**83.42/86.65**	**0.518**	**0.786**

TABLE II
COMPARISON ON MG-SIMS. MODELS WITH † ARE REPRODUCED UNDER THE SAME CONDITIONS. BEST RESULTS ARE MARKED IN BOLD.

Methods	Acc-7	F1	MAE	Corr
TFN†	74.30	75.62	0.782	0.615
LMF†	75.53	76.88	0.771	0.628
MulT†	72.94	73.66	0.803	0.594
MISA†	73.85	74.59	0.795	0.602
MAG-BERT†	70.75	71.75	0.842	0.559
Self-MM†	78.76	79.85	0.728	0.685
ALMT†	77.70	78.94	0.745	0.662
DEVA†	81.56	82.32	0.678	**0.752**
MAF-Net	**83.14**	**84.25**	**0.655**	0.748

effectively mitigates unimodal dominance, ensuring a balanced contribution from all modalities.

Moreover, it is worth noting that the scenarios in the MG-SIMS dataset are distinct due to their low-resource nature and linguistic specificity (Mongolian). Therefore, it is more challenging to model the multimodal data without large-scale pre-training. However, as shown in Table I, MAF-Net demonstrates competitive performance across nearly all metrics on both MOSI and MOSEI datasets. On MOSI, compared to the sub-optimal methods, MAF-Net achieves substantial improvements on Acc-5 and Acc-7, with gains of 2.88% and 3.65% over Self-MM and DEVA, respectively, while also attaining the lowest MAE of 0.702. On MOSEI, MAF-Net outperforms the best competing methods across all metrics, achieving an Acc-2 of 83.65/86.58, an Acc-7 of 55.12, and an F1-score of 83.42/86.65, surpassing ALMT and ConFEDE by notable margins. These consistent improvements across two widely used benchmarks demonstrate the strong generalization capability of MAF-Net in multimodal sentiment analysis.

TABLE III
ABLATION STUDY OF KEY COMPONENTS ON THE MOSI DATASET. GSA DENOTES THE GEOMETRIC SEMANTIC ALIGNMENT, AND MMIF DENOTES THE MODULAR MASKED INTERACTIVE FUSION. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	**Components**		**Metrics**		
	GSA	**MMIF**	**Acc-5 (%)**	**MAE (↓)**	**Corr (↑)**
Baseline	×	×	53.64	0.7040	**0.7935**
w/o MMIF	✓	×	54.81	0.7027	0.7902
w/o GSA	×	✓	53.94	0.7034	0.7905
MAF-Net	✓	✓	**55.10**	**0.7023**	0.7913

TABLE IV
ABLATION STUDY ON THE MODALITY ALIGNMENT MODULE. WE COMPARE DIFFERENT ALIGNMENT STRATEGIES TO VALIDATE THE EFFECTIVENESS OF OUR GEOMETRIC SEMANTIC ALIGNMENT (GSA).

Core Mechanism	**Acc-5 (%)**	**MAE (↓)**	**Corr (↑)**
w/o Alignment	54.66	0.7058	0.7903
w/ Simple Linear	53.21	0.7197	0.7858
w/ Cross-Attention	53.21	0.7224	0.7916
w/o Gating Ratio	53.50	0.7062	**0.7927**
GSA (Ours)	**55.10**	**0.7023**	0.7913

E. Ablation Study

To rigorously validate the effectiveness of the proposed framework and verify our hypotheses regarding semantic alignment and modality balance, we conduct a comprehensive ablation study on the MOSI dataset. We organize our analysis into three complementary perspectives: (1) the contribution of key architectural components, (2) the mechanisms underlying geometric semantic alignment, and (3) the efficacy of modular masked interaction in handling challenging scenarios.

1) Impact of Key Components: Table III shows that the full model performs best, indicating that GSA and MMIF are complementary. Removing GSA leads to a clear drop (Acc-5:

TABLE V
ABLATION ON TEXTUAL GUIDANCE SOURCES. WE COMPARE FIXED-LAYER ALIGNMENT (USING ONLY H_t^1, H_t^2, OR H_t^L) AGAINST OUR LAYER-WISE HIERARCHICAL STRATEGY ($H_t^i \rightarrow \text{GSA}^i$).

Guidance Source	Acc-5 (%)	MAE (↓)	Corr (↑)
Bottom-only	53.64	0.7140	0.7870
Middle-only	53.06	0.7177	0.7866
Top-only	53.64	0.7045	**0.7920**
GSA (Ours)	**54.81**	**0.7027**	0.7902

TABLE VI
ABLATION STUDY ON THE THREE COMPONENTS OF MODULAR MASKED INTERACTIVE FUSION. WE EVALUATE EACH COMPONENT'S CONTRIBUTION BY REMOVING IT FROM THE FULL MODEL.

Strategy	Acc-5 (%)	MAE (↓)	Corr (↑)
w/o $\mathcal{M}_{\text{pos}}$	53.64	0.7189	0.7853
w/o $\mathcal{M}_{\text{sync}}$	54.37	0.7187	0.7851
w/o $\mathcal{M}_{\text{emo}}$	52.19	0.7205	0.7881
MMIF (Ours)	**55.10**	**0.7023**	**0.7913**

$55.10 \rightarrow 53.94$), suggesting that explicit cross-modal semantic alignment is necessary for effective fusion. Removing MMIF also degrades performance (Acc-5: $55.10 \rightarrow 54.81$; MAE increases), supporting that regulating inter-modal information flow helps alleviate unimodal dominance and improves multi-source utilization.

2) Mechanism Analysis of Geometric Semantic Alignment: We compare alignment strategies in Table IV. Simple linear projection and cross-attention underperform our Gram-based alignment, indicating that matching global geometric structure is more effective than direct dimension-wise mapping or parameter-heavy token-wise interaction on MOSI. Removing the gating ratio further hurts performance, showing that adaptive alignment strength is important.

We further study textual guidance sources in Table V. Our layer-wise hierarchical guidance outperforms using any single fixed layer, suggesting that aligning non-verbal features with textual representations at multiple abstraction levels is beneficial.

Ablation Study on Modular Masked Interactive Fusion Components Table VI shows the ablation results for the three mask components. Removing $\mathcal{M}_{\text{emo}}$ causes the largest accuracy drop from 55.10% to 52.19% (-2.91%), while removing $\mathcal{M}_{\text{pos}}$ and $\mathcal{M}_{\text{sync}}$ leads to decreases of 1.46% and 0.73%, respectively. Consistent patterns are observed in MAE and correlation metrics, where the full model consistently outperforms all ablated variants.

V. CONCLUSION

In this paper, we presented MAF-Net, a novel framework designed to address the persistent challenges of cross-modal semantic gaps and unimodal dominance in Multimodal Sentiment Analysis. By integrating Geometric Semantic Alignment (GSA) to project heterogeneous non-verbal features into a unified textual manifold and employing Modular Masked Interactive Fusion (MMIF) to regulate inter-modal information flow, our approach effectively captures subtle emotional cues while mitigating noise and redundancy. Extensive experiments on standard benchmarks (MOSI, MOSEI) and our newly constructed Mongolian dataset (MG-SIMS) demonstrate that MAF-Net achieves state-of-the-art performance, validating its superior robustness in complex interaction scenarios and its strong potential for low-resource language applications.

ACKNOWLEDGE

This study is supported by the National Natural Science Foundation of China (62466044), Inner Mongolia Key Research and Achievement Transformation Plan Project (2025YFHH0115, 2025YFHH0215), and the "Young Science and Technology Talent Support Program" of Higher Education Institutions in Inner Mongolia Autonomous Region (NJYT23059).

REFERENCES

[1] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[2] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[3] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[5] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019, 2019, p. 6558.

[6] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[7] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 2359–2369.

[8] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[9] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably)," in *International conference on machine learning*. PMLR, 2022, pp. 9226–9259.

[10] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8238–8247.

[11] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "Pmr: Prototypical modal rebalance for multimodal learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 029–20 038.

[12] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, "Boosting multi-modal model performance with adaptive gradient modulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 214–22 224.

[13] Y. Wei, R. Feng, Z. Wang, and D. Hu, "Enhancing multimodal cooperation via sample-level modality valuation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 338–27 347.
[14] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Information Fusion*, vol. 95, pp. 306–325, 2023.
[15] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.
[16] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.
[17] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis," *arXiv preprint arXiv:2310.05804*, 2023.
[18] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[19] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.
[20] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
[21] S. Wu, D. He, X. Wang, L. Wang, and J. Dang, "Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 2, 2025, pp. 1601–1609.
[22] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2247–2256.
[23] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1642–1651.
[24] D. Yang, H. Kuang, S. Huang, and L. Zhang, "Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1708–1717. [Online]. Available: https://doi.org/10.1145/3503161.3547755
[25] J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu, "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)*, 2023, pp. 7617–7630.
[26] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "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*, 2021, pp. 2554–2562.