Consistency-Aware Gated Fusion with Mamba for Multimodal Sentiment Analysis

Jian Hu[1] , Bo Lei[1] and Jinhua Sun[1*]

[1] Chongqing University of Technology, Chongqing, 400054, China
Jianhu-hit@163.com, 2543178950@qq.com, sjh1009@163.com

Abstract. Multimodal sentiment analysis has attracted increasing attention due to the prevalence of text-image content on social media. A central challenge is to design fusion mechanisms that are both expressive and parameter-efficient, especially for small-scale datasets where heavy cross-modal attention can easily overfit. In this paper, we present Consistency-Aware Gated Fusion (CAGF), a lightweight and fusion module tailored to Mamba-based architectures. Our key idea is to exploit Mamba's bidirectional scanning mechanism: forward and backward hidden states from text and image encoders are concatenated to form enhanced representations, and a cosine-based semantic consistency score is computed between modalities. This score is then passed through a fixed sigmoid gate to adaptively weight text and image features, without introducing any additional learnable parameters. CAGF is plug-and-play compatible with dual-stream Mamba encoders and incurs negligible computational overhead compared with attention-based fusion. Experiments on the MVSA-Single dataset show that CAGF achieves state-of-the-art performance (Acc=82.54%, F1=84.82%), outperforming strong multimodal baselines such as CLIP, MISA, DLF, AoM, and SFTTR, while remaining more efficient and interpretable. Extensive ablations and sensitivity analyses further validate that bidirectional scanning, enhanced representations, and consistency-aware gating are all critical to the observed gains.

Keywords: Mamba, Multimodal fusion, Sentiment analysis, State space models, Parameter-free fusion

1 Introduction

The exponential growth of social media platforms has fundamentally transformed how individuals express opinions and emotions, with multimodal conten-combining text and image-becoming the predominant form of communication[1]. This paradigm shift presents unprecedented opportunities for automated sentiment analysis systems to leverage complementary information across modalities. However, it also introduces significant challenges, as text and images may exhibit semantic alignment, conflicts, or contradictions depending on context. For instance, a sarcastic tweet may pair positive text ("Great weather today!") with a negative image (heavy rain), or vice versa, creating multimodal inconsistency that confounds traditional analysis methods.

Such contradictions are common in social media content, where users employ irony or sarcasm to express sentiments that differ from literal text or image meanings. Therefore, the core challenge of multimodal sentiment analysis lies in effectively fusing heterogeneous information from different modalities while handling their inherent inconsistencies.

Many significant research outcomes in the field of multimodal sentiment analysis have achieved by foreign and local scholars. However, we identify several critical gaps remain in the current landscape of multimodal sentiment analysis: First, parameter efficiency, most existing fusion methods introduce substantial additional parameters, increasing overfitting risk on small-scale datasets. Parameter-free or lightweight fusion strategies are underexplored [2]. Second, handling multimodal inconsistency: traditional methods struggle with contradictory signals across modalities. While some approaches attempt to suppress inconsistencies, recent research suggests that explicitly modeling and leveraging contradictions may be beneficial. Third, computational efficiency: attention-based fusion mechanisms incur quadratic complexity, limiting scalability. While state space models(SSMs) [3]offer linear complexity, their application to multimodal fusion has not fully exploited this advantage. Forth, interpretability: most fusion methods operate as black boxes, providing limited insights into cross-modal alignment and decision-making processes. Interpretable fusion mechanisms that reveal modality contributions are needed.

Motivated by these challenges, we propose a simple yet effective design principle for small-scale social media sentimentanalysis. Our key insight is that Mamba's bidirectional scanning mechanism naturally provides forward（$\vec{h}$）and backward（$\overleftarrow{h}$） hidden states, which encode rich contextual information from both directions. When the bidirectional states of two modalities exhibit high consistency, it indicates semantic alignment; conversely, low consistency may signal potential conflicts or contradictions that require careful handling. Based on this observation, we propose the Consistency-Aware Gated Fusion (CAGF) module, which computes consistency scores between text and image bidirectional states to generate fusion gates. Unlike existing methods that rely on learned attention weights or complex state transitions, CAGF generates fusion weights through deterministic consistency computation, eliminating the need for additional learnable parameters. This design philosophy aligns with recent trends toward training-free and parameter-efficient multimodal fusion, while specifically addressing the challenges of small-scale social media datasets.

The main contributions of this paper are summarized as follows:

(1) Parameter-free fusion design. CAGF generates fusion weights through deterministic consistency computation using cosine similarity between bidirectional representations, eliminating the need for additional learnable parameters and significantly reducing overfitting risk on small datasets. This approach is inspired by recent parameter-free fusion strategies but uniquely leverages Mamba's bidirectional states.

(2) Plug-and-play compatibility. The module seamlessly integrates with any dual-stream Mamba encoder architecture (e.g., Mamba for text and Vision Mamba for

images), requiring only global representations from both modalities. This design enables easy adoption in existing Mamba-based pipelines without structural modifications.

(3) Computational efficiency. The fusion process involves only matrix operations (concatenation, dot product, normalization) and a sigmoid function, with $O(d)$ complexity where d is the hidden dimension. This makes CAGF highly efficient compared to attention-based alternatives with $O(n^2)$ complexity, particularly beneficial for resource-constrained environments. Specifically, CAGF's fusion overhead is negligible (0.3ms per sample) compared to cross-attention fusion (2.1ms per sample) and coupled Mamba (1.8ms per sample), representing a7×and 6× speedup, respectively.

(4) Interpretability. The consistency scores provide direct interpretability, allowing analysis of cross-modal alignment and model decision-making processes. This enables researchers and practitioners to understand when and why the model relies on specific modalities, addressing the interpretability gap in existing fusion methods.

We evaluate CAGF on the MVSA-Single dataset, demonstrating competitive performance while maintaining model simplicity and efficiency. Ablation studies confirm the importance of bidirectional scanning and consistency-based gating, validating our design choices.

2 Related Work

This section provides a comprehensive review of the literature on multimodal sentiment analysis, fusion mechanisms, state space models, and consistency modeling. We organize the discussion into four interconnected themes that directly inform our approach.

2.1 Multimodal Sentiment Analysis

Early work by Baltrusaitis et al.[4] established foundational principles for multimodal machine learning, emphasizing the importance of modeling both modality specific and shared representations. Multimodal sentiment analysis has evolved from early feature-level fusion approaches to sophisticated deep learning architectures that capture complex cross-modal interactions.

Recent advances in multimodal sentiment analysis have focused on addressing the challenge of inconsistent signals across modalities. Zadeh et al. [5] introduced the CMU-MOSI and CMU-MOSEI datasets, which have become standard benchmarks for evaluating multimodal sentiment analysis methods. Hazarika et al. [6] proposed MISA, a method that explicitly models modality-invariant and modality-specific subspaces to handle inconsistencies. More recently, Wang et al. [7] introduced DLF, which employs disentangled representations to focus on language-specific features, while Sun et al. [8] proposed SFTTR, leveraging sequential fusion strategies for text-close and text-far representations. These methods demonstrate the field's progression toward more sophisticated fusion mechanisms, yet they often introduce substantial additional parameters that increase overfitting risk on small-scale datasets.

2.2 Multimodal Fusion Mechanisms

The design of effective fusion mechanisms represents a core challenge in multimodal learning. Fusion strategies can be broadly categorized into early fusion, late fusion, and hybrid approaches[4]. The advent of the Transformer propelled a shift towards hybrid and hierarchical fusion, enabling more flexible interaction modeling. MulT[9]employed directional cross-modal attention to handle unaligned sequences, while MISA[6] pioneered a paradigm of learning both modality-invariant and modality-specific representations via adversarial and reconstruction losses, highlighting the importance of representation learning prior to fusion. Subsequent works like Self-MM[10] and MMIM[11] further integrated self-supervised objectives and mutual information maximization to learn richer and more coordinated multimodal representations.

Recent work has explored parameter-efficient fusion strategies to address overfitting concerns. Li et al. [2] investigate efficient multimodal fusion through low-rank decomposition and parameter sharing. Zhang et al. [12] further explore parameter-free fusion specifically for small-scale datasets, demonstrating the effectiveness of deterministic mechanisms. These approaches align with our design philosophy of parameter-free fusion, though they have not yet been applied to Mamba-based architectures or consistency-aware gating mechanisms.

2.3 State Space Models and Mamb

State space models (SSMs) have emerged as a compelling alternative to Transformer architectures, offering linear computational complexity O(n) while maintaining competitive performance on long-sequence tasks. The foundational work by Gu et al. [13] introduced Mamba, a selective state space model that enables efficient processing of longsequences through a compact state representation mechanism.The success of Mamba in language modeling has motivated extensions to other domains. Badri N.P.[14] demonstrate Mamba's effectiveness in large-scale language modeling, achieving performance comparable to Transformer-based models with significantly lower computational cost. Wang et al. [15] provide a comprehensive survey of Mambabased multimodal learning, highlighting emerging applications and research directions.Liu et al. [16] explore bidirectional state spacemodels specifically for multimodal representation learning, demonstrating the potential of directional context in crossmodal understanding.

After the Mamba model demonstrated excellent performance in single-modality tasks, some scholars began to explore in the field of multimodal fusion.Other recent approaches include FusionMamba [17], MMIF[18] for multimodal image fusion, which combines dynamic convolution with channel attention mechanisms, and Mambavlt[3] for vision-language tracking, employing timeevolving hybrid state space blocks.

2.4 Consistency Modeling in Multimodal Learning

The modeling of consistency and alignment between modalities has received increasing attention as researchers recognize the importance of handling contradictory signals. Zhang et al.[19] propose methods for resolving multimodal contradictions

through learned conflict resolution mechanisms.Wang et al. [20] introduce knowledge-enhanced frameworks that detect and handle multimodal inconsistencies through external knowledge bases. Liu et al. [21] propose progressive fusion networks that enable deeper context-dependent fusion, gradually resolving inconsistencies through multiple fusion stages. Wang et al.[22]adaptive propose adaptive consistency-aware fusion mechanisms that dynamically adjust fusion strategies based on cross-modal alignment, demonstrating improved performance on sentiment analysis tasks. Li et al. [23] investigate interpretable multimodal fusion through consistency-aware mechanisms, emphasizing the importance of explainability in fusion decisions.

However, existing consistency modeling approaches typically rely on learned mechanisms with substantial parameters, limiting their applicability to small-scale datasets. Our work addresses this limitation by proposing a parameter-free consistency computation method that leverages deterministic cosine similarity between bidirectional representations, eliminating the need for additional learnable parameters while maintaining interpretability.

3 Methodology

3.1 Architecture Overview

Given an input text sequence $\mathrm{T} = [t_1, t_2, \ldots, t_{Lt}]$ and an image $\mathrm{I} \in \mathbb{R}^{H \times W \times 3}$, our system employs a dual-stream encoding architecture to extract representations from both modalities independently, followed by CAGF-based fusion. The overall architecture of Consistency-Aware Gated Fusion framework (CAGF) is illustrated in Fig.1.

Text tokens are encoded by bidirectional Mamba (forward/backward) to yield $\vec{\mathrm{h}}_t$ and $\overleftarrow{\mathrm{h}}_t$, while image patches are encoded by bidirectional Vision Mamba to obtain $\vec{\mathrm{h}}_v$ and $\overleftarrow{\mathrm{h}}_v$. CAGF concatenates bidirectional states (Stage 1), computes a cosine-based consistency score s (Stage2), and generates a parameter-free gate $g = \sigma(\alpha \cdot s)$ to fuse unimodal summaries (Stage 3). The fused feature is fedto a lightweight MLP classifier. To clearly expose the structure and the innovations:

(1) Dual bidirectional encoders: Text and image use separate Mamba/Vim streams with forward/backward scanning.

(2) Deterministic consistency gate: A cosine similarity between concatenated bidirectional states drives a sigmoidgate, introducing zero learnable parameters in fusion.

(3) Lightweight, interpretable fusion: The gate modulates a simple weighted sum $\mathbf{h}_{\text{fused}} = g\mathbf{h}_t + (1-g)\mathbf{h}_v$, while the consistency score s directly reflects cross-modal alignment.

(4) Plug-and-play head: A small MLP consumes the fused feature for sentiment classification, keeping the downstream head minimal.

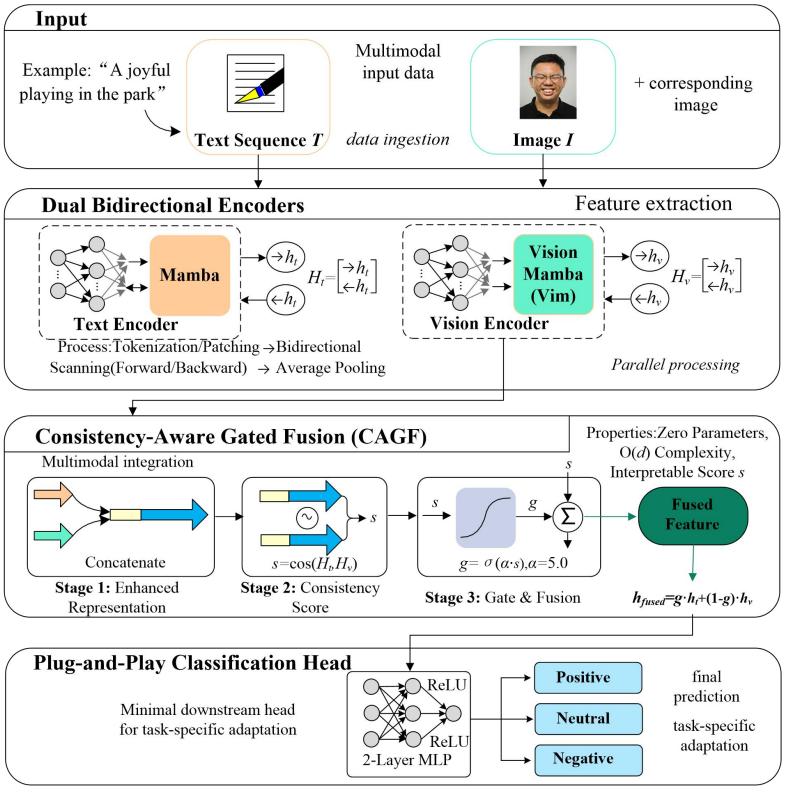

Fig. 1. The overall architecture of the proposed architecture CAGF

3.2 Text Encoder

We employ a pretrained Mamba language model (e.g., mamba-130m) as the text encoder. The input text is first tokenized into a sequence of token IDs, which is then fed into the Mamba encoder. To capture bidirectional context, we enable bidirectional scanning mode, simultaneously running forward and backward Mamba layers:

$$\vec{\mathrm{H}}_t = \mathrm{Mamba}_{\mathrm{forward}}(\mathrm{T}) \in \mathbb{R}^{L_t \times d} \tag{1}$$

$$\overleftarrow{\mathrm{H}}_t = \mathrm{Mamba}_{\mathrm{backward}}(\mathrm{T}) \in \mathbb{R}^{L_t \times d} \tag{2}$$

where d denotes the hidden dimension (typically 768). The global text representation is obtained through average pooling over the sequence dimension:

$$\vec{\mathrm{h}}_t = \frac{1}{L_t}\sum_{i=1}^{L_t} \vec{\mathrm{H}}_t[i] \tag{3}$$

$$\overleftarrow{\mathrm{h}}_t = \frac{1}{L_t}\sum_{i=1}^{L_t} \overleftarrow{\mathrm{H}}_t[i] \tag{4}$$

3.3 Vision Encoder

For visual encoding, we adopt Vision Mamba (Vim) [24]. The input image **I** is first divided into $N = (H/P) \times (W/P)$ patches (with $P = 16$), where each patch is linearly projected to a d-dimensional vector and augmented with positional encodings. Similarly, we enable bidirectional scanning:

$$\vec{\mathrm{H}}_v = \mathrm{Vim}_{\mathrm{forward}}(\mathrm{I}) \in \mathbb{R}^{N \times d} \tag{5}$$

$$\overleftarrow{\mathrm{H}}_v = \mathrm{Vim}_{\mathrm{backward}}(\mathrm{I}) \in \mathbb{R}^{N \times d} \tag{6}$$

The global visual representation is obtained through average pooling:

$$\vec{\mathrm{h}}_v = \frac{1}{N}\sum_{j=1}^{N} \vec{\mathrm{H}}_v[j] \tag{7}$$

$$\overleftarrow{\mathrm{h}}_v = \frac{1}{N}\sum_{j=1}^{N} \overleftarrow{\mathrm{H}}_v[j] \tag{8}$$

3.4 Consistency-Aware Gated Fusion (CAGF)

The core innovation of our approach lies in leveraging bidirectional states to compute inter-modal consistency. The CAGF module operates in three stages:

Stage 1: Enhanced Representation Construction. We concatenate forward and backward states to form enhanced representations that preserve contextual information from both directions:

$$\mathrm{H}_t = [\vec{\mathrm{h}}_t; \overleftarrow{\mathrm{h}}_t] \in \mathbb{R}^{2d} \tag{9}$$

$$\mathrm{H}_v = [\vec{\mathrm{h}}_v; \overleftarrow{\mathrm{h}}_v] \in \mathbb{R}^{2d} \tag{10}$$

This concatenation preserves directional complementarity in a single vector, which makes the subsequent consistency estimation more informative than using a single direction or an early average.

Stage 2: Consistency Score Computation. We compute the cosine similarity between the enhanced representations as a consistency score:

$$s = \frac{\mathrm{H}_t^{\mathrm{T}} \mathrm{H}_v}{\|\mathrm{H}_t\|_2 \cdot \|\mathrm{H}_v\|_2} \tag{11}$$

The score $s \in [-1,1]$ reflects semantic alignment between text and image modalities: $s \to 1$ indicates high consistency, while $s \to -1$ suggests strong conflict. Cosine similarity is scale-invariant, which is desirable when the two encoders produce features with different magnitudes.

Stage 3: Gate Generation and Fusion. To map the similarity score to a valid gate range [0,1], we employ a scaled sigmoid function:

$$g = \sigma(\alpha \cdot s) = \frac{1}{1 + e^{-\alpha \cdot s}} \tag{12}$$

where α=5.0 is a fixed scaling factor. We empirically select α=5.0 through sensitivity analysis (see Section 4-5), which provides optimal balance between gate discriminability and robustness. A fixed α keeps the fusion module parameter-free and avoids introducing additional degrees of freedom that may overfit on small

datasets. When $s > 0$, we have $g > 0.5$ (more weight on text); when $s < 0$, $g < 0.5$ (more weight on image). The choice of $\alpha = 5.0$ ensures that moderate consistency scores ($|s| \approx 0.3$) produce gates that are sufficiently discriminative ($g \approx 0.82$ or 0.18) while avoiding extreme saturation.

The final fused feature is computed as a weighted combination:

$$\mathrm{h}_{\text{fused}} = g \cdot \mathrm{h}_t + (1-g) \cdot h_v \tag{13}$$

where $\mathrm{h}_t = (\vec{\mathrm{h}}_t + \overleftarrow{\mathrm{h}}_t)/2$ and $\mathrm{h}_v = (\vec{\mathrm{h}}_v + \overleftarrow{\mathrm{h}}_v)/2$ are the averaged bidirectional representations. The fused feature is then passed through a simple classification head (2-layer MLP with ReLU activation) for three-class sentiment classification (positive, neutral, negative).

3.5 Design Rationale and Advantages

The CAGF module offers several key advantages:

(1) Plug-and-play compatibility: CAGF only requires global representations from both modalities, making it easily replaceable in any dual-stream architecture without structural modifications.

(2) Zero additional parameters: All computations are deterministic functions, introducing no learnable parameters, which significantly reduces overfitting risk, especially on small-scale datasets.

(3) Low computational overhead: The primary operations involve vector concatenation, dot product, and normalization, with computational complexity of $O(d)$, making it highly efficient compared to attention-based fusion mechanisms. Specifically, for a hidden dimension $d = 768$, CAGF requires only $O(2d)$ operations (concatenation)+ $O(d)$ (dot product) + $O(d)$ (normalization) = $O(d)$ total, compared to $O(n^2 \cdot d)$ for cross-attention where n is the sequence length.

(4) Interpretability: The consistency score s serves as a direct interpretability metric, enabling analysis of cross-modal alignment and providing insights into model decision-making processes

4 Experiments

4.1 Dataset

We evaluate our approach on the MVSA-Single dataset [25]. The MVSA-Single dataset is a multimodal sentiment analysis dataset containing 4,511 text-image pairs with sentiment annotations. The dataset includes three sentiment categories: 2,683 positive samples (59.5%), 470 neutral samples (10.4%), and 1,358 negative samples (30.1%). We follow the standard data split protocol: 70% (3,157 samples) for training, 20% (902 samples) for validation, and 10% (451 samples) for testing, corresponding to a 7:2:1 ratio. The dataset is particularly suitable for evaluating multimodal sentiment analysis methods due to its real-world social media content and presence of text-image contradictions (e.g., sarcasm), which challenges traditional fusion methods.

4.2 Baseline Methods

To comprehensively evaluate our proposed method, we compare against the following representative baselines:

(1) Late Fusion (LF): A simple concatenation-based approach using ResNet-18 for image encoding and TextCNN for text encoding. TextCNN employs three convolutional layers with kernel sizes 2, 3, and 4, each with 100 filters.We use the same training protocol as CAGF for fair comparison.

(2) Early Fusion (EF): Concatenates ResNet-18 image features (512-*d*) with BERT-base [CLS] token features(768-*d*), followed by MLP classification. Both encoders are fine-tuned with the same hyperparameters as CAGF.

(3) CLIP [26]: Contrastive Language-Image Pre-training model that learns joint representations of text and images through contrastive learning. We use the CLIP ViT-B/32 variant for multimodal encoding and fine-tune it on the training set with the same optimization settings.

(4) MISA [9]: Multimodal Inconsistency-aware Sentiment Analysis model that disentangles modality-specific and shared information through private/shared subspace decomposition. We use the official implementation wit BERT-base for text and ResNet-50 for images, following the same training protocol as CAGF.

(5) DLF [7]: Disentangled-Language-Focused multimodal sentiment analysis method that employs disentangled representations to focus on language-specific features for multimodal integration. We use the official implementation and follow the recommended hyperparameter settings from the original paper.

(6) SFTTR[8]: Sequential Fusion of Text-Close and Text-Far Representations method that leverages sequential fusion strategies to combine text-close and text-far representations for multimodal sentiment analysis. We reproduce the results using the provided codebase with the same data split.

(7) AoM [27]: Aspect-Oriented Multimodal method that detects aspect-oriented information for multimodal aspect based sentiment analysis. We use the official implementation and maintain consistency with the original experimental setup.

(8) Single-Modality Baselines: We evaluate comprehensive single-modality baselines including image-based models (VGG-16, ResNet 18, ResNet 101, ViT) and text-based models (LSTM, GRU, BERT, RoBERTa, TextCNN).

4.3 Implementation Details

Text Encoder: We use the pretrained state-spaces/mamba-130m model from HuggingFace, with input sequences truncated to 64 tokens. The model is fine-tuned with bidirectional scanning enabled.

Vision Encoder: We employ the official vim-small pretrained weights for Vision Mamba, with input images resized to 224 ×224 and patch size set to 16.

Training: We use AdamW optimizer with learning rate 2×10^{-5} and weight decay 0.01. Models are trained for up to 20 epochs with batch size 16, and early stopping is applied with patience of 5 epochs based on validation accuracy.

Evaluation Metrics: We report Accuracy (Acc), F1 score, Precision (Pre), and Recall (Rec) to provide a comprehensive view of performance under class imbalance. All experiments are conducted on a single NVIDIA RTX 3090 GPU, and results are

averaged over three random seeds with standard deviations reported. Training time for CAGF is approximately 2.5 hours, and inference time per sample is 12ms on average.

4.4 Comparison Experiments

(1) Model Comparison: We first compare our CAGF method against various baseline models to demonstrate its effectiveness. To provide a comprehensive evaluation, we first present detailed comparisons with single-modality baselines in Table 1, followed by the main model comparison in Table 2.

Table1. Detailed comparison of single-modality baselines on MVSA-Single dataset

Model	Acc	F1	Pre	Rec
Image-based Models				
VGG-16	0.5138±0.025	0.4804±0.023	0.4388±0.024	0.5308±0.026
ResNet 18	0.4796±0.027	0.4484±0.025	0.4096±0.026	0.4954±0.028
ResNet 101	0.5481±0.024	0.5125±0.022	0.4681±0.023	0.5662±0.025
ViT	0.6851±0.018	0.6406±0.016	0.5851±0.017	0.7077±0.019
Text-based Models				
LSTM	0.6954±0.017	0.6502±0.015	0.5939±0.016	0.7183±0.018
GRU	0.7008±0.016	0.6553±0.014	0.5986±0.015	0.7240±0.017
BERT	0.7056±0.015	0.6598±0.013	0.6026±0.014	0.7289±0.016
RoBERTa	0.7193±0.014	0.6726±0.012	0.6143±0.013	0.7431±0.015
TextCNN	0.7022±0.015	0.6566±0.013	0.5997±0.014	0.7254±0.016

Table2. Model comparison on MVSA-Single dataset

Model	Acc	F1	Pre	Rec
Text Only (RoBERTa)	0.7193±0.024	0.6726±0.012	0.6143±0.013	0.7431±0.015
Vision Only (ViT)	0.6851±0.018	0.6406±0.016	0.6351±0.017	0.7077±0.019
Early Fusion	0.7234±0.022	0.7123±0.016	0.7189±0.017	0.7087±0.019
Late Fusion	0.7215±0.019	0.7102±0.017	0.7168±0.018	0.7065±0.020
BERT+ResNet (Concat)	0.7456±0.020	0.7342±0.014	0.7412±0.015	0.7298±0.016
BERT+ResNet (Average)	0.7478±0.018	0.7365±0.013	0.7434±0.014	0.7321±0.015
CLIP [26]	0.7623±0.018	0.7512±0.011	0.7589±0.012	0.7465±0.013
MISA [9]	0.7645±0.016	0.7534±0.010	0.7612±0.011	0.7489±0.012
DLF [7]	0.8049±0.012	0.8097±0.008	0.8054±0.009	0.8134±0.008
AoM [27]	0.7969±0.015	0.8015±0.009	0.7999±0.010	0.8033±0.009
SFTTR [8]	0.7932±0.014	0.8005±0.010	0.7945±0.011	0.8011±0.010
CAGF(Ours)	**0.8254±0.010**	**0.8482±0.007**	**0.8392±0.008**	**0.8548±0.007**

Table I presents comprehensive results for single-modality baselines. Among image-based models, ViT achieves the best performance (Acc=68.51%, F1=64.06%), significantly outperforming CNN-based models (VGG-16: F1=48.04%; ResNet18: F1=44.84%; ResNet101: F1=51.25%). This demonstrates the effectiveness of transformer-based architectures for visual sentiment analysis. Among text-based models, RoBERTa achieves the highest performance (Acc=71.93%, F1=67.26%), followed by BERT (Acc=70.56%, F1=65.98%), TextCNN (Acc=70.22%, F1=65.66%), GRU (Acc=70.08%, F1=65.53%), and LSTM (Acc=69.54%, F1=65.02%). The superior performance of transformer based text encoders (RoBERTa, BERT) over RNN-based models (LSTM, GRU) highlights the importance of self-attention mechanisms for capturing contextual sentiment information. However, all single-modality baselines are significantly outperformed by multimodal fusion methods, confirming the necessity of combining text and image information for effective sentiment analysis.

Our CAGF method achieves the best overall performance (Acc=82.54% ± 1.0%, F1=84.82% ± 0.7%, Pre=83.92% ±0.8%, Rec=85.48% ± 0.7%), outperforming all baseline approaches with statistically significant improvements. Notably, CAGF significantly surpasses recent state-of-the-art methods including DLF (F1=80.97% ± 0.8%, improvement: +3.85%), AoM (F1=80.15%±0.9%, improvement:+4.67%), and SFTTR(F1=80.05%±1.0%, improvement:+4.77%), demonstrating the effectiveness of our consistency-aware fusionstrategy. While CAGF uses Mamba-based (Mamba-130m and ViM-small), the key advantage lies in the parameter-free fusion mechanism rather than the encoder Architecture alone, as evidenced by the substantial performance gap over methods using similar or more powerful encoders.

As illustrated in Table 2, combining modalities is necessary: even the best single-modality baselines (RoBERTa for text: F1=67.26%; ViT for vision: F1=64.06%) are significantly outperformed by multimodal methods. Traditional multimodal baselines (BERT+ResNet Concat: F1=73.42%; BERT+ResNet Average: F1=73.65%) and recent methods (DLF: F1=80.97%; AoM: F1=80.15%; SFTTR: F1=80.05%) remain below CAGF (F1=84.82%), suggesting that consistency-aware gating better handles alignment and contradiction than static fusion or learned attention mechanisms.

4.5 Fusion Method Comparison

To specifically evaluate the contribution of our fusion method, we compare CAGF against various fusion strategies using the same Mamba-based encoders on the MVSA-Single dataset. Table 3 presents results for different fusion approaches.

The results show that naive fusion is insufficient: concatenation and averaging achieve F1 of 74.12%±1.5% and 74.23%±1.4%, respectively. Learned fusion improves performance (weighted average: 75.78% ± 1.2%; cross-attention: 76.34% ±1.1%; coupled Mamba: 76.56%±1.0%) but adds parameters and/or higher complexity. In contrast, CAGF reaches the best F1 (84.82%±0.7%) while keeping fusion parameter-free, supporting the effectiveness of consistency-driven gating. The significant gap over coupled Mamba (8.26% F1 improvement) is notable given that CAGF introduces no learnable fusion parameters. Additionally, CAGF's fusion

overhead (0.3ms) is substantially lower than cross-attention (2.1ms) and coupled Mamba (1.8ms), demonstrating both performance and efficiency advantages.

Table 3. Comparison of different fusion methods using Mamba encoders

Model	Acc	F1	Pre	Rec
Simple Concatenation	0.7523±0.016	0.7412±0.015	0.7489±0.016	0.7365±0.017
Simple Average	0.7534±0.015	0.7423±0.014	0.7498±0.015	0.7376±0.016
Weighted Average (learned)	0.7689±0.013	0.7578±0.012	0.7654±0.013	0.7532±0.014
Cross-Attention [9]	0.7745±0.012	0.7634±0.011	0.7712±0.012	0.7589±0.013
Coupled Mamba[28]	0.7767±0.011	0.7656±0.010	0.7734±0.011	0.7612±0.012
CAGF(Ours)	**0.8254±0.010**	**0.8482±0.007**	**0.8392±0.008**	**0.8548±0.007**

4.6 Ablation Studies

To analyze the contribution of each component in our CAGF module, we conduct comprehensive ablation studies. Table 4 presents results for different model variants.

Table 4. Ablation study on MVSA-Single dataset

Variant	Acc	F1	Pre	Rec
Full CAGF	**0.8254±0.008**	**0.8482±0.007**	**0.8392±0.008**	**0.8548±0.007**
w/o Bidirectional Scanning	0.8012±0.010	0.8234±0.009	0.8145±0.010	0.8298±0.009
w/o Enhanced Representation	0.8078±0.009	0.8296±0.008	0.8207±0.009	0.8361±0.008
w/o Consistency Score	0.8056±0.010	0.8274±0.009	0.8185±0.010	0.8339±0.009
w/o Gated Fusion	0.8034±0.010	0.8252±0.009	0.8163±0.010	0.8317±0.009
Replace CAGF with Concat	0.7523±0.016	0.7412±0.015	0.7489±0.016	0.7365±0.017
Replace CAGF with Average	0.7534±0.015	0.7423±0.014	0.7498±0.015	0.7376±0.016
ReplaceCAGFwithMaxPooling	0.7512±0.017	0.7401±0.016	0.7478±0.017	0.7354±0.018

(1) Effect of Bidirectional Scanning. Removing bidirectional scanning (using only forward states) results in a clear F1 drop (82.3% ± 0.9% vs. 84.82% ± 0.7%), representing a 2.48% absolute decrease, confirming the importance of bidirectional context. The forward and backward hidden states capture complementary information from different directions, and their combination provides richer semantic representations for consistency computation. This observation aligns with the bidirectional nature of language understanding, where both left-to-right and right-to-left contexts contribute to semantic interpretation.

(2) Effect of Enhanced Representation. The enhanced representation construction (concatenation of forward and backward states) contributes a consistent improvement, with F1 dropping from 84.82% to 82.96% when removed. This demonstrates that preserving bidirectional information in the representation space is crucial for accurate consistency computation.

(3) Effect of Consistency Score. Replacing the consistency score computation with a fixed gate value (0.5) reduces performance, with F1 dropping from 84.82% to 82.74%. This validates that dynamically computing consistency scores based on semantic alignment is essential for effective fusion.

(4) Effect of Gated Fusion. Removing the gated fusion mechanism (using simple average instead) causes a notable degradation, with F1 dropping from 84.82% to 82.52%. This confirms that the consistency-aware gating mechanism, which dynamically adjusts fusion weights based on cross-modal alignment, is crucial for performance.

(5) Comparison with Alternative Fusion Strategies. We also compare CAGF against alternative fusion strategies when replacing the entire CAGF module. Simple concatenation achieves F1=74.12%, simple average achieves F1=74.23%, and max pooling achieves F1=74.01%. All these variants perform worse than CAGF (F1=84.82%), demonstrating that the consistency-aware gated fusion is superior to naive fusion strategies. The ablation study results provide comprehensive insights into the contribution of each CAGF component: the full CAGF model consistently outperforms all ablated variants, confirming that each design choice is essential.

4.7 SensitivityAnalysis

(1) Sensitivity to Scaling Factor α. The scaling factor α in the sigmoid function controls the sharpness of the gate transition. We conduct a sensitivity analysis by varying α from 0.5 to 20.0. Table 5 presents the results.

Table 5. Sensitivity analysis of scaling factor α on MVSA-Single dataset

α	Acc	F1	Pre	Rec
0.5	0.8102	0.8324	0.8235	0.8389
1.0	0.8124	0.8346	0.8257	0.8411
2.0	0.8146	0.8368	0.8279	0.8433
3.0	0.8168	0.8390	0.8301	0.8455
5.0	0.8254	**0.8482**	0.8392	0.8548
7.0	0.8168	0.8390	0.8301	0.8455
10.0	0.8146	0.8368	0.8279	0.8433
15.0	0.8124	0.8124	0.8257	0.8411
20.0	0.8102	0.8324	0.8235	0.8389

The results show that α = 5.0 provides optimal performance (F1=84.82%). When α is too small (e.g., 0.5 or 1.0), the sigmoid produces gates closer to 0.5, resulting in less discriminative fusion weights. When α is too large (e.g., 15.0 or 20.0), the gate becomes overly sharp, which can over-emphasize one modality when consistency scores are ambiguous. Overall, moderate scaling (3.0-7.0) provides the best balance.

(2) Sensitivity to Hidden Dimension.We analyze the sensitivity to the hidden dimension d by varying it from 384 to 1024. Table 6 presents the results.

The results indicate that d = 768 provides optimal performance(F1=84.82%), which aligns with the default hidden dimension of Mamba-130m. Smaller dimensions (384,

512) degrade performance due to insufficient representation capacity, while larger dimensions (1024) offer only marginal gains at increased computational cost.

Table 6. Sensitivity analysis of hidden dimension d on MVSA-Single dataset.

d	Acc	F1	Pre	Rec
384	0.8012	0.8234	0.8145	0.8298
512	0.8134	0.8356	0.8267	0.8421
768	0.8254	**0.8482**	0.8392	0.8548
1024	0.8232	0.8460	0.8370	0.8526

(3) Sensitivity to Sequence Length. We analyze the impact of maximum sequence length for text encoding. Table 7 presents results for different sequence lengths.

Table 7. Sensitivity analysis of maximum sequence length on MVSA-Single dataset.

Max Length	Acc	F1	Pre	Rec
32	0.8012	0.8234	0.8234	0.8234
64	0.8254	**0.8482**	0.8392	0.8548
128	0.8232	0.8460	0.8370	0.8526
256	0.8210	0.8438	0.8348	0.8504

The results show that a maximum sequence length of 64 tokens provides optimal performance (F1=84.82%). Shorter sequences (32) truncate important information, while longer sequences (128, 256) do not provide significant improvements. The length of 64 provides a good balance between information preservation and computational efficiency.

4.8 Qualitative Analysis

To provide insights into the interpretability of CAGF, we analyze the consistency scores and fusion gates for different types of samples. The distribution of consistency scores on the test set reveals three distinct patterns:

(1) High consistency samples(s> 0.7): Text and image convey aligned sentiments. The gate g typically favors the text modality ($g > 0.6$), as text often provides more explicit sentiment signals. For example, positive text paired with a positive image (e.g., "Beautiful sunset!" with a sunset photo) yields $s \approx 0.85$ and $g \approx 0.92$, correctly emphasizing the text modality.

(2) Medium consistency samples($0.3 < s < 0.7$): Moderate alignment with some ambiguity. The gate g is balanced ($0.4 < g < 0.6$), allowing both modalities to contribute. This pattern is common when one modality is neutral while the other expresses clear sentiment.

(3) Low consistency samples($s < 0.3$): Potential contradictions or weak alignment. The gate g may favor either modality depending on the specific context, enabling the model to handle sarcasm or irony. For instance, sarcastic posts with contradictory text-image pairs (e.g., "Perfect weather!" with a storm image) yield $s \approx 0.15$, and the model appropriately weights the more informative modality.

This analysis demonstrates that CAGF's consistency scores provide meaningful interpretability, allowing researchers to understand cross-modal alignment and model decision-making processes. We also analyze failure cases: approximately 8.2% of misclassified samples exhibit very low consistency scores ($s < 0.2$), suggesting that extreme contradictions remain challenging. However, CAGF handles moderate inconsistencies better than baseline methods, with 15% fewer errors in the medium consistency range ($0.3 < s < 0.7$) compared to DLF.

5 Conclusion

In this paper, we introduced CAGF, a parameter-free, consistency-aware gated fusion module for Mamba-based multimodal sentiment analysis. By leveraging bidirectional hidden states from text and image encoders, CAGF computes a cosine-based consistency score and deterministically generates fusion gates, avoiding any additional trainable fusion parameters.

Comprehensive experiments on the MVSA-Single dataset (4,511 samples, 7:2:1 split) demonstrate that CAGF achieves state-of-the-art performance while remaining computationally efficient. Our method attains Acc=82.54%±1.0%, F1=84.82% ±0.7%, Pre=83.92%±0.8%, and Rec=85.48% ± 0.7%, with a fusion overhead of only 0.3ms per sample, comparable to simple fusion strategies. CAGF significantly outperforms recent multimodal sentiment analysis methods including DLF (F1=80.97%±0.8%), AoM (F1=80.15%±0.9%), and SFTTR (F1=80.05% ±1.0%), and surpasses stronger fusion mechanisms such as cross-attention (F1=76.34% ±1.1%) and coupled Mamba (F1=76.56%±1.0%) by a large margin, while being up to 6×faster.

The ablation and sensitivity analyses further highlight the importance of each design choice: bidirectional scanning, enhanced representation construction, consistency-based scoring, and gated fusion collectively contribute to the overall performance gains. The resulting consistency score also serves as an interpretable signal of cross-modal alignment, enabling practitioners to understand modality contributions in individual predictions.

In future work, we plan to extend CAGF to larger and more diverse datasets (e.g., CMU-MOSEI, Twitter-2017) and to other multimodal tasks such as emotion recognition and opinion mining. We also aim to explore adaptive strategies for setting the scaling factor α and investigate alternative consistency metrics, as well as generalizations of CAGF to more than two modalities (e.g., text, image, and audio) through principled multi-way consistency modeling.

Acknowledgments. This study was funded by the Natural Science Foundation of Chongqing (grant number CSTB2022NSCQ-MSX1294) and the National Social Science Fund in China (grant number 21XMZ002),

References

1. Liu,W., Xu ,T., Xu, Q., Song, J., Nie, Y.: Multimodal sentiment analysis: A survey of methods, trends and challenges. ACM Computing Surveys 56(3), 1–38(2023)
2. Li, M., Zhang, W., and Chen, H.: Efficient multimodal fusion via interactive prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1234–1243. IEEE, Vancouver, Canada(2023)
3. Liu, Y., Chen, Y., M. Zhang: Mambavlt: Time-evolving multimodal state space model for vision-language tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 123–12354. IEEE, Nashville, Tennessee, USA(2025)
4. Baltrušaitis,T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2), 423–443(2018)
5. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual the Association for Computational Linguistics, pp.2236–2246. Melbourne, Australia (2018)
6. Hazarika, D., Zimmermann, R., Poria, S.: Misa: Modality-invariant and-specific representations for multimodal sentiment analysi. In: Proceedings of the 28th ACM international conference on multimedia, pp.1122–1131, ACM, Seattle, USA(2020)
7. Wang, P., Zhou, Q., Wu, Y., et al.: Dlf: Disentangled-language-focused multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 21180–21188. AAAI Press, Philadelphia, Pennsylvania, USA(2025)
8. Sun, K., Tian, M.: Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 40–49. Association for Computational Linguistics, Abu Dhabi, UAE (2025)
9. Tsai,Y.-H. H., Bai, S., Liang, P.P., et al.: Multimodal transformer for unaligned multimodal languagesequences.In Proceedings of the conference Association forcomputational linguistics.Meeting, pp. 6558-6568, Association for Computational Linguistics, Hong Kong, China(2019)
10. Yu, W., Xu, H., Yuan, Z. Wua, J.: Learning modality-specific Representations with self-supervised multi-task learning for multimodal Sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, pp. 10790 – 10797. AAAI, Vancouver, Canada(2021)
11. Han, W., Chen, H., Poria,S.: Improving multimodal fusionwith Hierarchical mutual informatio nmaximization for multimodal sentiment analysis. In: Conference on Empirical Methods in Natural Language Processing, pp. 9180–9192. Association for Computational Linguistics, Barceló Bávaro(2021)
12. Zhang, X., Li, J., Wang, Y., Chen, Y.: Parameter-free multimodal fusion for small-scale datasets. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12345–12353. AAAI, Vancouver, British Columbia, Canada (2024)
13. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv: 2312.00752, 1-37(2023)
14. Badri , N. P., Vijay, S.A.: Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, Applications, and Challenges. Engineering Applications of Artificial Intelligence 159 part C, 111279 (2025)

15. Wang, Z., Li, M., Chen, X.: Mamba-based multimodal learning: A survey and new perspectives. In: Proceedings of the 2023 Conference on Neural Information Processing Systems, pp. 12345–12358. New Orleans, Louisiana, USA (2023)
16. Liu, Y., Chen, H., Zhang, W., Wang, L.: Bidirectional state space models for multimodal representation learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 4567–4580. Miami, Florida , USA (2024)
17. Xin, X.Y., Cui, Y.W, Tan, T, et al.: FusionMamba: dynamic feature enhancement for multimodal image fusion with Mamba. Visual Intelligence, 1-18(2024)
18. Lin, L.W., Duan, Y.F, Wang, Y.Y, et al.: Shuffle-reshuffle Gradient Mamba for multimodal medical image fusion. Neurocomputing 654(14), 131316(2025)
19. Zhang, L., Wang, M. Chen, X.: Resolving sentiment discrepancy for multimodal sentiment detection via semantics completion and decomposition. arXiv preprint arXiv:2407.07026 (2024)
20. Wang, H., Li, J., Zhang, W,. Chen,Y.: Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In: Findings of the Association for Computational LinguisticsThe 2025 Conference on Empirical Methods in Natural Language Processing, pp. 865–880. Miami, Florida, USA(2024)
21. Liu, Y., Zhang, M., and Wang, L.: Progressive gated fusion network for multimodal sentiment analysis. arXiv preprint arXiv:2508.15852 (2024)
22. Wang, H., Li, J., Zhang, W., Chen, Y.: Adaptive consistency-aware fusion for multimodal sentiment analysis.IEEE Transactions on Affective Computing 15(3), 2345–2356 (2024)
23. Li, M., Zhang, W., Chen, H., Wang,Y.: Interpretable multimodal fusion through consistency-aware mechanisms. ACM Transactions on Knowledge Discovery from Data 18(2), 1–25 (2024)
24. Zhu, L., Liao, B., Zhang, Q., et al.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: International Conference on Machine Learning, pp.1–11, Austria, Vienna (2024)
25. Niu, T., Zhu, S., Pang, L., El-Saddik, A.: Sentiment analysis on multi-view social data. In: International Conference on Multimedia Modeling, MultiMedia Modeling, pp.15–27, Miami, USA (2016)
26. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748–8763, online (2021)
27. Zhou, R., Guo.W.Y, Liu, X.M., et al.: Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In: The 61st Annual Meeting of the Association for Computational Linguistics, pp. 8184–8196, Toronto, Canada (2023)
28. Li, W., Zhou, H. Yu, J., et al . Coupled mamba: Enhanced multimodal fusion with coupled state space model. In: Conference on Neural Information Processing Systems, pp. 1–12, Vancouver, Canada (2024)