Enhancing Multimodal Sentiment Analysis through the Integration of Attention Mechanisms and Spiking Neural Networks

Anonymous ACL submission

Abstract

The success of Vision Transformers has sparked growing interest in integrating the self-attention mechanism and Transformer-based architecture into Spiking Neural Networks (SNNs), aiming to combine the brain-inspired efficiency of SNN with the power of attention-based models. While recent efforts have introduced spiking-compatible self-attention modules, they often suffer from two key limitations: the absence of effective scaling strategies and architectural bottlenecks that hinder the extraction of fine-grained local features and the integration of multimodal information. To address these issues, we introduce the Spiking-Generated Multimodal Transformer, which features a spiking self-attention mechanism with biologically plausible and computationally efficient scaling. Unlike conventional spiking models that focus narrowly on single modalities or shallow representations, our model adopts a multi-stage architecture, including both single-modal processing and modality fusion networks, enabling a deeper understanding and integration of complex multimodal inputs like audio, text, and visual signals. This synergistic design allows the model to leverage the temporal dynamics of spikes while maintaining high-level semantic alignment across modalities. As a result, our approach improves both energy efficiency and performance. Experiments on benchmark datasets, including SIMS and MOSEI for multimodal sentiment analysis, validate the effectiveness of our approach.

1 Introduction

Spiking Neural Networks (SNNs), inspired by biological neural systems, are considered the third generation of artificial neural networks(Maass, 1997). In recent years, SNNs have been successfully integrated with various deep learning architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers, which has shown its promise in tasks like image analysis (Lan et al., 2023; Patel et al., 2021), robotics(Lele et al., 2020; Rueckert et al., 2016) and sequence modeling.

Unlike traditional networks, SNNs exhibit discontinuous and temporally dynamic behaviors. They excel in modeling chaotic, such as Lorenz systems, where a single spike can significantly alter subsequent spike sequences (Nicola and Clopath, 2017). Moreover, SNNs operate across diverse dynamical states, subcritical, critical, supercritical and periodic, by adjusting neuron parameters (Liang and Zhou, 2022). These characteristics make SNNs a compelling choice for complex modeling tasks.

Multimodal sentiment analysis integrates text, audio, and visual inputs for sentiment prediction. There are many datasets(Yu et al., 2020; Zadeh et al., 2018b) and related studies(Liu et al., 2018; Han et al., 2021; Sun et al., 2023) about this issue. A key challenge is aligning and integrating modalities effectively. Inspired by the brain's architecture, we model each modality as input to interconnected SNN neurons and apply a shared SNN-based structure for consistent single-modal feature extraction prior to fusion.

Neuroscience research suggests that Transformers equipped with recurrent positional mimic the spatial representation of the hippocampal formation(Whittington et al., 2021). Motivated by this connection, we propose integrating SNNs with Transformers to leverage their complementary strengths in modeling temporal and sequential information. Our contributions are as follows:

- We propose an SNN-based fusion framework that combines multi-layer integration and Transformer architectures for effective multimodal sentiment analysis.
- We exploit the synergy between SNN dynamics and Transformer modeling to advance multimodal affective computing.

- We conduct extensive experiments, demonstrating our model's superiority over existing approaches.

2 Related Work

2.1 Spiking Neural Networks

Spiking Neural Networks (SNNs), regarded as the third generation of artificial neural networks, are biologically inspired and known for their energy efficiency and suitability for temporal processing due to their spike-based, event-driven nature (Maass, 1997). Recent research has explored integrating SNNs with deep learning, particularly Transformers, inspired by findings that recurrent positional encoding in Transformers resembles hippocampal spatial representations (Whittington et al., 2021).

Spikformer (Zhou et al., 2022) is the first directly trained spiking Vision Transformer, introducing a spiking self-attention mechanism by activating the Query, Key, and Value with spiking neurons and replacing softmax with spike-based neurons. It also replaces Transformer components like layer normalization and GELU activation with batch normalization and spiking neurons. Based on this, Spikingformer (Zhou et al., 2023) achieves a purely spike-driven architecture by modifying the residual connections. Spike-driven Transformer (Yao et al., 2023) further reduces energy consumption by proposing a linear-complexity spike-driven attention mechanism. However, these models rely on shallow convolutional layers to extract local features and form patch sequences, and they lack effective scaling strategies for input vectors.

2.2 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) aims to predict sentiment by integrating textual, acoustic, and visual modalities. Traditional modality fusion strategies are typically categorized into early fusion (which combines raw features) and late fusion (which combines model outputs) (Lu et al., 2023), yet both struggle to effectively capture cross-modal dependencies. To address this, deep learning methods have introduced more sophisticated fusion techniques, including attention-based mechanisms (Yuan et al., 2021) and Graph Neural Networks (GNNs) for modeling inter-modal relationships (Zeng et al., 2023). However, the majority of these approaches are still built upon conventional deep-learning architectures.

Recent studies have explored biologically inspired fusion strategies, but the application of SNNs in MSA remains largely unexplored. Given their spike-driven, temporally dynamic nature, SNNs offer a compelling alternative for multimodal integration. In this work, we propose an SNN-based fusion framework, leveraging multi-layer feature integration and hybrid SNN-Transformer architecture to enhance sentiment classification.

3 Methodology

In this section, we provide a detailed description of the proposed method. We begin by defining the multimodal sentiment analysis task and its corresponding notation. Next, we introduce the overall model architecture of this article, with a particular focus on the spatial module. Finally, we outline the multi-task learning strategy and the overall optimization objective.

3.1 Task definition

The MSA task in this paper refers to predicting the polarity and intensity of sentiment through video information. MSA task usually contains three main modalities: text (denoted by T), audio (denoted by A,) and visual (denoted by V). We define the input as $\mathbf{X}_i \in \mathbb{R}^{T_i \times D_i}$, where T_i is the sequence length of modality i, D_i denotes the feature dimension of modality i, $i \in t, a, v$. We expect the model to integrate information from all modalities and assign an emotion score to the person in the video, representing both the polarity (indicated by the sign) and intensity (indicated by the absolute value).

3.2 Overall Architecture

The overall architecture of our model is illustrated in Figure 1, which mainly consists of three components: the multimodal data preprocessing module, the single-modal network, and the fusion network. The preprocessing module is responsible for transforming the raw multimodal inputs into coarse-grained features. However, due to the shallow structure of this module, it can only capture limited local information and generates only single-modal feature tensors. Therefore, we introduce a spatial module for further processing. This module includes both the single-modal network and the fusion network, which perform refined processing for each modality and fusion across modalities, respectively. Within these modules, we incorporate SNNs and attention mechanisms to encode the coarse features into a high-level semantic representation.

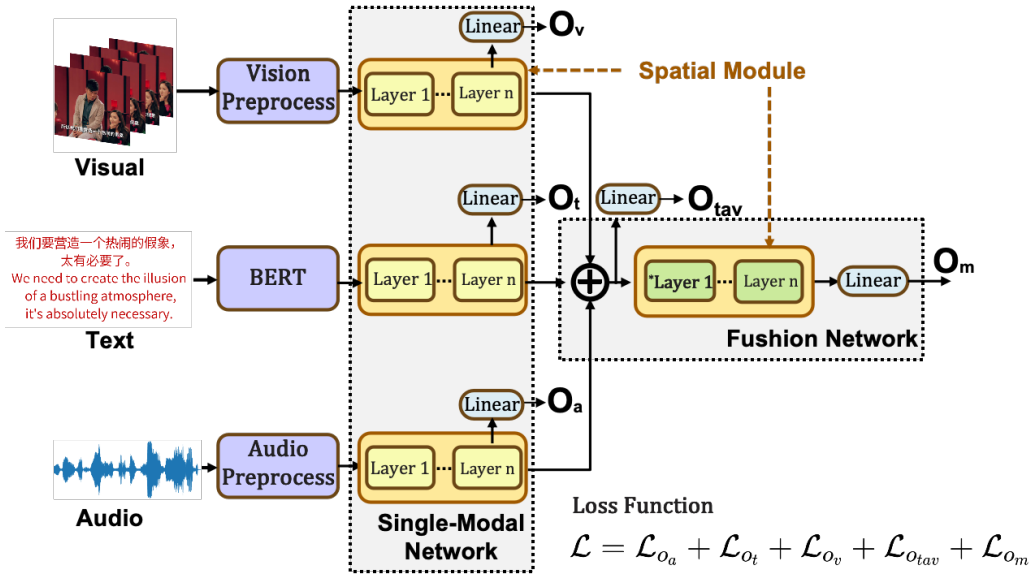


Figure 1: The overall architecture of our method. The O_a, O_t, O_v, O_{tav}, and O_m are the prediction outputs of the three single-modal and multi-modal tasks, respectively. The model components include a Single-Modal Module and a Fusion Network.

In previous architectures that combine SNNs with attention mechanisms, such as SpikingResFormer, the input data is primarily visual, resulting in relatively short temporal lengths T. However, in multimodal video-based emotion recognition tasks, T is typically much larger. Directly applying such architectures in this context leads to severe gradient vanishing problems. To address the limitations of these existing models in handling multimodal emotion recognition, while retaining the strengths of SNN-attention integration, we propose a novel architecture named **Spiking-Generated Multimodal Transformer**, which combines the residual structure of Transformer with a biologically inspired spiking self-attention mechanism.

Each spatial layer consists of two modules: the (Connectted) Multi-Head Spike-generated Self-Attention (($\mathcal{C}$)MHSGSA) block and a Spiking Feed-Forward Network with residual. Finally, the model ends with a classification layer to produce the final result. Additionally, to facilitate multi-task training, we attach separate classifiers to the outputs of each modality prior to the fusion network.

3.3 Spatial Module

As illustrated in Figure 1, the spatial module consists of a Multi-Head Spike-Generated Self-Attention (MHSGSA) module and a Spike-Generated Feed-Forward Network (SGFFN). We first introduce the two modules and then derive the form of the basic layers.

LIF Model. The LIF model (leaky integrate-and-fire) we used here is a computationally simplified version of the more biologically meaningful conductance-based LIF model. The specific version used in this study is shown below, which is consistent with the formulations adopted in previous works that integrate SNNs and Transformer:

$$\begin{aligned} U_i[t] &= V_i[t] + \frac{1}{\tau}(I_i[t] - (V_i[t] - V_{rest})) \\ s_i[t] &= H(U_i[t] - V_{th}) \\ V_i[t+1] &= s_i[t]V_{rest} + (1 - s_i[t])U_i[t] \end{aligned} \tag{1}$$

where $H(\cdot)$ is the Heaviside function. The first equation represents the charge and leak processes, the second equation represents that the neuron will fire a spike when the neuron's potential reaches the threshold potential V_{th}, and the third equation represents that the neuron's potential will be reset to the resting potential V_{rest} after firing. The simplification process and detailed analysis can be found in Appendix A.

In subsequent sections, for an input current $\mathbf{I} \in \mathbb{R}^{T\times F}$, the above model can be used to calculate the output spike results $\mathbf{S} \in \{0,1\}^{T\times F}$, which we will denote as $\mathbf{S} = \text{LIF}(\mathbf{I})$ for simplicity.

Since the LIF model is a discontinuous model, for the elements in S that only take the values 0 and 1, we use a method similar to previous work(Shi et al., 2024) to approximate the gradient calculation with $arctan$ function.

Vanilla Self-Attention. We first review the classic Vanilla Self-Attention and propose our Spike-Generated Self-Attention model.

In the original Transformer paper(Vaswani et al., 2017), Vanilla Self-Attention is formulated as follows: for the input $\mathbf{X} \in \mathbb{R}^{n \times d}$,

$$\begin{aligned} \mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \mathbf{K} = \mathbf{X}\mathbf{W}_K, \mathbf{V} = \mathbf{X}\mathbf{W}_V \\ VSA(\mathbf{X}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \end{aligned} \tag{2}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_m \times d}$ are learnable weight matrices. In some papers(Zhou et al., 2022; Yao et al., 2023), the authors argue that introducing floating-point multiplication and exponential operations in softmax during the VSA process does not conform to the computational rules of SNNs. However, In some more recent papers that combine Transformers with SNNs(Shi et al., 2024), the authors added a convolution layer after the spike output; similarly, in(Nicola and Clopath, 2017), a double exponential filter is directly applied to the spike output to compute the firing rate for use in subsequent layers, indicating that introducing floating-point operations in spike output processing is reasonable. Therefore, our model will more fully retain the original architecture of VSA.

Spike-generated Self-Attention. We present Spike-Generated Self-Attention (SGSA), a novel mechanism tailored for spiking neural networks, enabling self-attention to integrate while efficiently managing multimodal feature representations.

In this module, due to the characteristics of the LIF model, it is necessary to control the magnitude of the input values. Therefore, we apply a normalization method to normalize the inputs. To accommodate sequential input, we adopt Layer Normalization as the normalization function. A scaling factor $scale_1$ is then applied to control the data magnitude. The specific value of this factor is discussed in the experimental section; in short, it is currently treated as a tunable hyperparameter. The normalized and scaled input is then passed into the LIF model to obtain the SNN output of this module.

$$\begin{aligned} \mathbf{X}_{norm} &= scale_1 LN(\mathbf{X}) \\ \mathbf{X}_{spike} &= LIF(\mathbf{X}_{norm}) \end{aligned} \tag{3}$$

Next, similar to Vanilla Self-Attention (VSA), we compute the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$.

$$\begin{aligned} \mathbf{Q} &= \mathbf{X}_{spike}\mathbf{W}_Q \\ \mathbf{K} &= \mathbf{X}_{spike}\mathbf{W}_K \\ \mathbf{V} &= \mathbf{X}_{spike}\mathbf{W}_V \end{aligned} \tag{4}$$

Additionally, it is necessary to determine an appropriate scaling factor $scale_2$ to ensure numerical stability during the softmax operation. Inspired by the original VSA paper(Vaswani et al., 2017), for $\mathbf{Q}$ and $\mathbf{K}$ whose elements are independent and identically distributed (i.i.d.), with zero mean and unit variance, the dot product $\mathbf{Q}\mathbf{K}^T$ yields a matrix where each element has a mean of 0 and variance of d, where d is the dimensionality of the key/query vectors. Thus, dividing by $\sqrt{d}$ can standardize the result.

Our model applies a similar idea to compute $scale_2 = \sqrt{\sigma_1^2\sigma_2^2 d d_m^2 f^2}$. Here f denotes the average firing rate of all neurons over the entire time interval, i.e., $\frac{\text{Total number of spikes during this period}}{\text{Total time length} \times \text{Number of neurons}}$, and $\sigma_1^2\sigma_2^2$ is the product of the variances of the matrices $\mathbf{W}_Q$ and $\mathbf{W}_K$. The detailed derivation can be found in Appendix B. Based on this scaling, the remaining computation of the Spike-Generated Self-Attention (SGSA) is formulated as follows:

$$SGSA(\mathbf{X}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{scale_2}\right)\mathbf{V} \tag{5}$$

Here, we propose the single-head form of SGSA. It can be easily extended to the multi-head SGSA (MHSGSA) following a similar approach to the vanilla Transformer(Vaswani et al., 2017). In MHSGSA, we employ d parallel SGSA attention modules to process the input, and then concatenate their outputs as shown below:

$$MHSGSA(\mathbf{X}) = \begin{bmatrix} SGSA_1(\mathbf{X}) \\ SGSA_2(\mathbf{X}) \\ \vdots \\ SGSA_d(\mathbf{X}) \end{bmatrix} \tag{6}$$

Connected Spike-generated Self-Attention. In the MHSGSA module, we do not introduce connections between neurons, as this module mainly serves to preprocess individual modalities. Each neuron only needs to respond to the information it receives. However, during the multimodal fusion process, it is also necessary for neurons to integrate information from different modalities. Therefore, in the multimodal fusion part, we introduce the $\mathcal{C}$MHSGSA module: similar to the approach

in (Nicola and Clopath, 2017), we incorporate a non-trainable connection matrix generated from a normal distribution $\omega \in \mathbb{R}^{d_m \times d_m}$ into the original LIF model as follows:

$$U_i[t] = V_i[t] + \frac{1}{\tau}(I_i[t] - (V_i[t] - V_{rest}))$$

where the input current no longer consists solely of external input, but also includes interactions between neurons. Specifically, the equation can be rewritten as:

$$U_i[t] = V_i[t] + \frac{1}{\tau}(I_i[t] + \sum_{j=1}^{d_m} \omega_{ij} s_j[t] - (V_i[t] - V_{rest}))$$

If we denote the output of this new LIF model as $\mathbf{S} = LIF'(\mathbf{X})$, then replacing all instances of LIF in MHSGSA with LIF' yields the $\mathcal{C}$MHSGSA module.

Single-modal Network and Fusion Network. Following the architecture of MHSGSA and $\mathcal{C}$MHSGSA modules, we define two key components of our model: the Single-Modal Network and the Fusion Network, each responsible for distinct aspects of representation learning.

The Single-Modal Network is designed to process single-modal inputs and its basic layer is structured by attaching a Spiking Feed-Forward Network with residual connection is attached after MHSGSA as follows:

$$\begin{aligned} \mathbf{Y}_{norm} &= scale_1 LN(MHSGSA(\mathbf{X})) \\ \mathbf{Y}_{spike} &= LIF(\mathbf{Y}_{norm}) \\ \mathbf{Y}_{out} &= linear(\mathbf{Y}_{spike}) \\ \mathbf{R}_{norm} &= scale_1 LN(\mathbf{Y}_{out}) \\ \mathbf{R} &= \mathbf{R}_{norm} + \mathbf{X}_{norm} + \mathbf{Y}_{norm} \end{aligned} \quad , \quad (7)$$

where $\mathbf{R}$ represents the output of the corresponding module. The Fusion Network incorporates a cross-modal attention mechanism and a variant of the spiking activation unit, enabling effective integration of multi-modal information. Its structure is largely similar to that of the Single-Modal Network, with the primary difference being the substitution of the LIF model by its connected counterpart.

Both networks adopt a residual connection before the SNN module to alleviate the vanishing gradient problem during training. While their processing pipelines are structurally aligned, the key difference lies in their attention strategies: the Single-Modal Network uses standard multi-head attention (MHSGSA) for single-modal data, whereas the Fusion Network employs cross-modal multi-head attention ($\mathcal{C}$MHSGSA) to support inter-modal feature fusion. In addition, the spiking function LIF$'$ is adapted in the Fusion Network to handle multi-modal signal dynamics better. Concatenating the outputs of these two networks yields the final multi-modal representation used by the model.

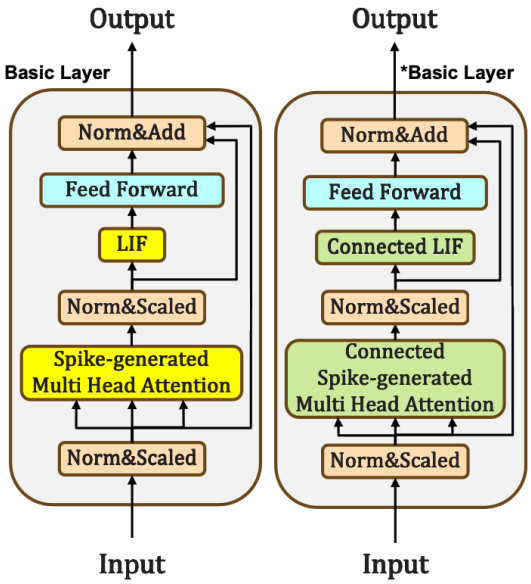


Figure 2: Architecture of the basic layer that makes up the Single-modal Network and the Fusion Network. Including Multi-Head Spike-generated Self-Attention (MHSGSA) and Connected Multi-Head Spike-generated Self-Attention ($\mathcal{C}$MHSGSA).

What's more, as stated in (Shi et al., 2024), in multimodal sentiment analysis, video, audio, and text information are closely related, yet methods based on single modality or direct feature fusion often struggle with fine-grained feature extraction and prediction accuracy. By treating single-modal and fused-modal predictions as sub-tasks and applying multi-task learning to share representations, the complementary features of different modalities can be better utilized, thereby improving the model's generalization performance. Moreover, compared with the fusion of multiple models, multi-task learning enables independent and joint optimization within a single model, significantly reducing computational complexity. Accordingly, we introduce our total loss as:

$$\mathcal{L}_{loss} = \lambda_a \mathcal{L}_a + \lambda_v \mathcal{L}_v + \lambda_t \mathcal{L}_t + \lambda_{tav} \mathcal{L}_{tav} + \lambda_m \mathcal{L}_m$$

The loss function $\mathcal{L}$ is computed using the MAE, with weighting coefficients of each component in the overall loss λ_a, λ_v, λ_t, λ_{tav} and λ_m set to 1, 1, 1, 3 and 6, respectively.

4 Experiments

In this section, we introduce the datasets used in the experiment and evaluate the performance of our model on the multimodal sentiment analysis task. Then, we perform ablation experiments on key components in our model. The experiment involves three modalities: video, text, and audio.

4.1 Dataset

CH-SIMSv2: The dataset CH-SIMS (Yu et al., 2020) contains 2281 carefully edited Chinese video clips from 60 film and television videos. Compared with the original dataset, the CH-SIMS v2.0(Liu et al., 2022) doubles its size with another 2121 refined video segments with both single-modal and multimodal annotations. The voice part of the video is in Mandarin, the length of the short clip is no less than 1 second and no more than 10 seconds. The annotation task includes 2 categories (Positive and negative), 3 categories (Positive, Negative, Neutral), and 5 categories (Negative, weakly Negative, neutral, weakly positive, and Positive). This dataset has accurate multimodal and independent single-modal annotations, which can be used to support researchers in multimodal or single-modal sentiment analysis. **MOSEI:** The Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) (Zadeh et al., 2018b) dataset is designed to enhance the diversity of training samples, including a broader range of topics and richer annotations. It contains 23,453 annotated video clips collected from online video-sharing platforms, involving 1,000 different speakers across 250 topics. The dataset provides sentiment scores on a 7-point Likert scale ranging from -3 (highly negative) to $+3$ (highly positive) for each sample.

4.2 Implementation Details

In this article, our model training is based on the NVIDIA V100 GPU. We use the Adam optimizer in the experiment to optimize the model parameters. To make a fair comparison with other baselines, we follow the preprocessing method of the previous work(Cheng et al., 2023; Jin et al., 2023; Cai et al., 2025) to extract features of images, audios, and texts. For visual features, we use Facet(Ekman and Rosenberg, 1997) to extract facial expression features from MOSEI, and OpenFace 2.0(Tadas et al., 2018) for the SIMS dataset. Both tools detect facial landmarks, action units (20 dimensions), head pose, and eye gaze. For audio, Librosa(McFee et al., 2015) extracts 33-dimensional acoustic features from SIMS (log-F0, 20 MFCCs, 12 CQT) at 22,050 Hz, while COVAREP(Degottex et al., 2014) is used for MOSEI, capturing MFCCs, pitch, glottal parameters, and prosody-related features. For text, we adopt the BERT pre-trained model(Devlin et al., 2019) to extract contextualized word embeddings, without requiring prior word segmentation. Finally, all modal features are organized as tensors with shape `[batch size, sequence length, feature dimensions]`.

4.3 Baselines and Metrics

The details of the baselines are as follows: **EF-LSTM:** utilizes feature-level fusion and sequence learning of Bidirectional Long-Short Term Memory (Bi-LSTM) deep neural networks.(Williams et al., 2018b) **LF-DNN:** introduces three model structures to encode multimodal data and then combines PCA for early feature fusion and late decision fusion.(Williams et al., 2018a) **TFN:** introduces an end-to-end fusion method for sentiment analysis by modeling the single-modal, bimodal, and trimodal interactions using a 3-fold Cartesian product from modality embeddings.(Zadeh et al., 2017) **LMF:** utilizes low-rank factors for multimodal representation and making multimodal feature fusion more efficient.(Liu et al., 2018) **MFN:** proposes a Delta-Memory Attention Network (DMAN) to identify cross-view interactions while summarizing through Multi-view Gated Memory.(Zadeh et al., 2018a) **Graph-MFN:** introduces the Dynamic Fusion Graph (DFG) module based on the MFN network and performs fusion analysis.(Zadeh et al., 2018b) **MulT:** utilizes a bidirectional cross-modal attention mechanism to focus on the interactions between multi-modal data at different time steps.(Tsai et al., 2019) **MLF-DNN:** is a multi-task version of LF-DNN.(Yu et al., 2020) **MLMF:** is a multi-task version of LMF.(Yu et al., 2020)

For metrics, on MOSEI, following previous works(Peng et al., 2023; Lin and Hu, 2023; Yuan et al., 2021), our evaluation indicators include binary accuracy (Acc-2), seven-class accuracy (Acc-7), F1-score, mean absolute error (MAE), and Pearson correlation coefficient (Corr). It is worth noting that Acc-2 has two different representation methods, namely negative/non-negative and negative/positive(Zadeh et al., 2018b). Here, we only use the negative/positive classification criterion. On SIMS, following previous works(Yu et al., 2020; Liang et al., 2021), our evaluation indicators

include Acc-2, Acc-3, Acc-5, F1-score, MAE, and Corr. The detailed formulations are provided in the Appendix C.

4.4 Main Results

The performance of our proposed model is presented in Table 1 and Table 2, alongside reproduced results from several competitive baseline models (Cai et al., 2025; Mao et al., 2022). Our model consistently surpasses all baselines in terms of Acc-5 and Acc-7, respectively. Beyond these primary metrics, the SNN-Transformer also achieves highly competitive or improved results across most other evaluation indicators, with only marginal gaps in a few cases. These results highlight the effectiveness of integrating biologically inspired spiking neural dynamics with Transformer-based architectures for multimodal sentiment analysis. The temporal encoding and energy-efficient event-driven processing of SNNs, combined with the sequence modeling capacity of Transformers, enable our model to better capture cross-modal dependencies and temporal nuances, leading to more robust and accurate sentiment classification.

Model	Acc-5	Acc-3	F1	Corr	MAE
ef_lstm	49.26	72.38	79.03	0.6588	0.3374
lf_dnn	52.76	72.59	79.46	0.7120	0.3014
tfn	52.55	72.21	80.14	0.7073	0.3031
lmf	47.79	64.90	73.88	0.5569	0.3672
mfn	54.53	73.66	81.19	0.7266	0.2954
graph_mfn	45.78	67.18	72.60	0.5743	0.3787
mult	54.81	73.19	80.73	0.7378	0.2905
mlf_dnn	49.67	68.47	76.62	0.6395	0.3352
mlmf	51.22	69.98	77.30	0.6703	0.3222
ours	**55.03**	73.21	80.46	0.7247	0.3006

Table 1: Performance comparison between our method and baselines on SIMS Dataset.

4.5 Ablation Study

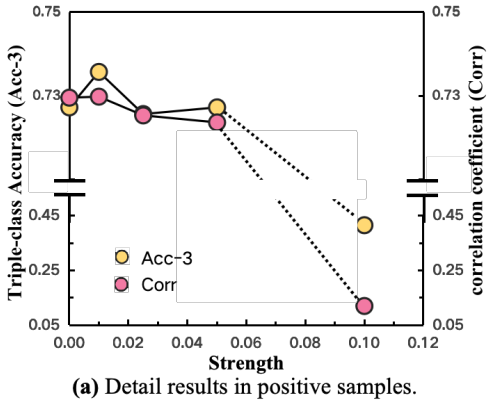
We conduct ablation studies on the proposed model, comprising two key parts: First, we investigate the effect of two critical hyperparameters, $scale_1$, and the inter-neuron connection strength within $\mathcal{C}$MHSGSA, on the stability and performance of the SNN. Second, we evaluate the model's ability to utilize multi-modal information by systematically reducing the number of input modalities.

Connected Couple Strength and Scaling Factors. In our experiments, the ω in CMHSGSA is generated from a normal distribution with mean 0 and variance 1 (corresponding to the case where the connection probability is 1 in (Nicola and Clopath, 2017)), and then scaled by a factor called $strength$, which is used to control the magnitude of the coupling strength. We performed a grid search over both $scale_1$ and $strength$, and the detailed results can be found in Appendix D. The best results were obtained when $scale_1 = 0.5$ and $strength = 0.01$.

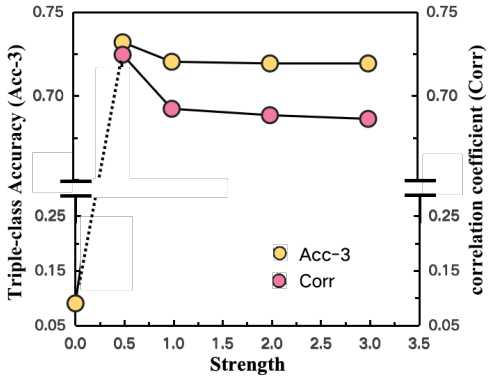
We plotted line charts for the corresponding row

Model	Acc-7	Acc-2	F1	MAE
ef_lstm	49.3	80.3	81.0	0.603
lf_dnn	52.1	82.3	82.2	0.561
tfn	50.2	82.5	82.1	0.593
lmf	48.0	82.0	82.1	0.623
mfn	51.3	82.8	82.8	0.573
graph_mfn	51.9	84.0	83.8	0.569
mult	51.8	82.5	82.3	0.580
mmim	51.9	83.8	83.6	0.599
ours	**52.2**	83.5	83.4	0.569

Table 2: Performance comparison between our method and baselines on MOSEI Dataset.

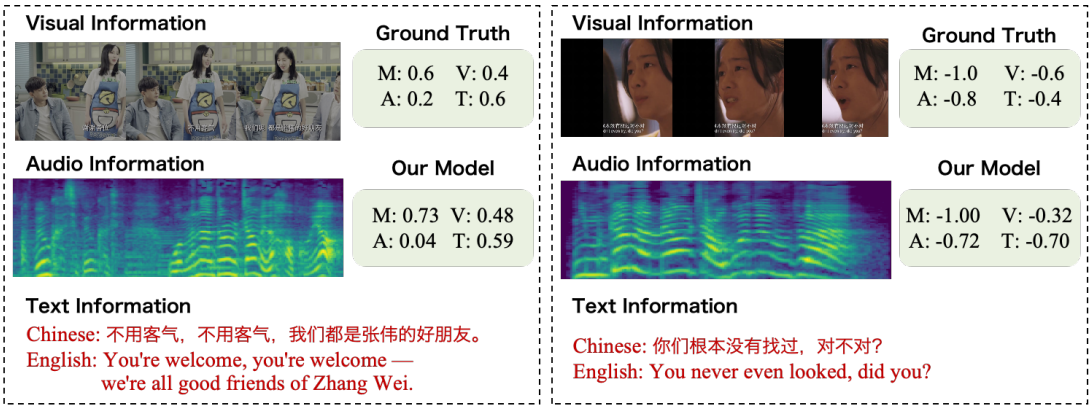


(a) Detail results in positive samples.



(b) Detail results in positive samples.

Figure 3: Detail results of our method in four test samples of SIMS dataset. left is the strength sample, and right is the scale curve.



(a) Detail results in positive sample. **(b)** Detail results in negative sample.

Figure 4: Detail results of our method in two test samples of SIMS dataset. (a) a sample with a positive sentiment type, and (b) a sample with a negative sentiment type.

and column of the best result, as shown in Figure 3. The figure shows that the model maintains strong performance across a relatively wide range of values. However, once these parameters exceed this range, the model's performance deteriorates sharply, which aligns with the inherent requirement for dynamical stability in spiking neural networks.

Multimodal Feature. We slightly modify the model architecture to accept inputs from only two modalities. Specifically, we remove the spatial module corresponding to the excluded modality and, before feeding into the fusion network, only merge the results from the remaining two modalities. In the loss function, we remove the loss component associated with the excluded modality; all other aspects remain unchanged. Using the optimal hyperparameter settings, we conduct experiments by removing each of the Text (T), Audio (A), and Visual (V) modalities individually. The results show that utilizing all three modalities outperforms using any combination of two modalities, indicating that our fusion network module effectively leverages information from all modalities.

4.6 Case Study

To better illustrate the performance of our model, we selected representative samples from the SIMS test set and examined the sentiment predictions for each modality in detail. The SIMS dataset was chosen because it provides ground-truth sentiment labels for each modality, allowing for a more comprehensive and modality-specific evaluation. Figure 4 shows these samples with the sentiment types of positive and negative.

Our analysis reveals that the model performs consistently well across different modalities. In all cases except when the true label is exactly zero, the predicted sentiment polarity (positive or negative) matches the ground-truth label, demonstrating that our multi-task training strategy effectively enhances the model's performance. More details about the neutral and cases with inconsistent sentiment across modalities (e.g., one modality being positive while another is negative) can be found in Appendix E.

modal	Acc-5	Acc-3	Acc-2	F1	MAE
A+V	48.55	66.83	73.60	73.51	0.3872
T+V	50.77	73.02	80.56	80.68	0.3209
T+A	50.97	67.99	75.24	75.32	0.3317
T+A+V	**55.03**	**73.21**	80.56	80.68	**0.3006**

Table 3: The result of multimodal feature ablation study on MOSEI Dataset.

5 Conclusion

We propose a novel class of techniques that combine Spiking Neural Networks with the self-attention mechanism, referred to as (Connected) Multi-Head Spike-Generated Self-Attention (($\mathcal{C}$)MHSGSA), which are tailored for single-modal processing and multimodal fusion. To enhance stability and control expressiveness, we introduce input and output scaling strategies. Based on this, we develop the Single-Modal Network and the Fusion Network, which are further unified into a Spiking-Generated Multimodal Transformer. Extensive experiments validate the effectiveness of our approach, achieving strong performance on multimodal sentiment analysis tasks.

Limitations

Despite its promising performance, our SNN-Transformer framework also presents certain limitations. First, training spiking neural networks remains computationally challenging due to the non-differentiable nature of spike events, requiring surrogate gradient methods that may introduce approximation errors. Second, achieving optimal results may require manual design of the neuron connection matrix; in particular, the relationship between the hyperparameters $scale_1$ and $strength$ necessary for proper model function remains insufficiently explored. Third, the development of more advanced fusion mechanisms and modality-adaptive spiking encoders is needed. Extending the framework to effectively handle noisy or asynchronous multimodal data remains an open and important challenge.

References

Yujian Cai, Xingguang Li, Yingyu Zhang, Jinsong Li, Fazheng Zhu, and Lin Rao. 2025. Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. *Scientific Reports*, 15(1):2126.

Hongju Cheng, Zizhen Yang, Xiaoqi Zhang, and Yang Yang. 2023. Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion. *IEEE Transactions on Affective Computing*, 14(4):3149–3163.

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)*, pages 960–964. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 6–15.

Cong Jin, Cong Luo, Ming Yan, Guangzhe Zhao, Guixuan Zhang, and Shuwu Zhang. 2023. Weakening the dominant role of text: Cmosi dataset and multimodal semantic enhancement network. *IEEE Transactions on Neural Networks and Learning Systems*.

Yuxiang Lan, Yachao Zhang, Xu Ma, Yanyun Qu, and Yun Fu. 2023. Efficient converted spiking neural network for 3d and 2d classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9211–9220.

Ashwin Sanjay Lele, Yan Fang, Justin Ting, and Arijit Raychowdhury. 2020. Learning to walk: Spike based reinforcement learning for hexapod robot central pattern generation. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 208–212. IEEE.

Junhao Liang and Changsong Zhou. 2022. Criticality enhances the multilevel reliability of stimulus responses in cortical neural networks. *PLoS computational biology*, 18(1):e1009848.

Xinyan Liang, Yuhua Qian, Qian Guo, Honghong Cheng, and Jiye Liang. 2021. Af: An association-based fusion method for multi-modal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9236–9254.

Ronghao Lin and Haifeng Hu. 2023. Multi-task momentum distillation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258.

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

Qiang Lu, Xia Sun, Yunfei Long, Zhizezhang Gao, Jun Feng, and Tao Sun. 2023. Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*.

Wolfgang Maass. 1997. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-sena: An integrated platform for multimodal sentiment analysis. *arXiv preprint arXiv:2203.12441*.

B McFee, C Raffel, D Liang, DPW Ellis, M McVicar, E Battenberg, and O Nieto. 2015. librosa: Audio and

music signal analysis in python no. scipy. In *Proceedings of the 14th Python in Science Conference*, pages 18–24.

Wilten Nicola and Claudia Clopath. 2017. Supervised learning in spiking neural networks with force training. *Nature communications*, 8(1):2208.

Kinjal Patel, Eric Hunsberger, Sean Batir, and Chris Eliasmith. 2021. A spiking neural network for image segmentation. *arXiv preprint arXiv:2106.08921*.

Junjie Peng, Ting Wu, Wenqiang Zhang, Feng Cheng, Shuhua Tan, Fen Yi, and Yansong Huang. 2023. A fine-grained modal label-based multi-stage network for multimodal sentiment analysis. *Expert Systems with Applications*, 221:119721.

Elmar Rueckert, David Kappel, Daniel Tanneberg, Dejan Pecevski, and Jan Peters. 2016. Recurrent spiking networks solve planning tasks. *Scientific reports*, 6(1):21142.

Xinyu Shi, Zecheng Hao, and Zhaofei Yu. 2024. Spikingresformer: bridging resnet and vision transformer in spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619.

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*, 15(1):309–325.

Baltrusaitis Tadas, Zadeh Amir, Lim Yao Chong, and Morency Louis-Philippe. 2018. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

James CR Whittington, Joseph Warren, and Timothy EJ Behrens. 2021. Relating transformers to models and neural representations of the hippocampal formation. *arXiv preprint arXiv:2112.04035*.

Jennifer Williams, Ramona Comanescu, Oana Radu, and Leimin Tian. 2018a. Dnn multimodal fusion techniques for predicting video sentiment. In *Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML)*, pages 64–72.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018b. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19.

Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. 2023. Spike-driven transformer. *Advances in neural information processing systems*, 36:64043–64058.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4400–4407.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

Yufei Zeng, Zhixin Li, Zhenjun Tang, Zhenbin Chen, and Huifang Ma. 2023. Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis. *Expert Systems with Applications*, 213:119240.

Chenlin Zhou, Liutao Yu, Zhaokun Zhou, Zhengyu Ma, Han Zhang, Huihui Zhou, and Yonghong Tian. 2023. Spikingformer: Spike-driven residual learning for transformer-based spiking neural network. *arXiv preprint arXiv:2304.11954*.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. 2022. Spikformer: When spiking neural network meets transformer. *arXiv preprint arXiv:2209.15425*.

A Simplification of the Conductance-Based LIF Model

The conductance-based Leaky Integrate-and-Fire (LIF) model is a more biologically interpretable model. Multi-channel versions of this model have been used in previous studies. In this work, we adopt the single-channel version:

$$\frac{dV}{dt} = \frac{V_{rest} - V}{\tau} + (V_{rev} - V)gG,$$

where V_{rest} represents the resting potential, and τ is the membrane time constant. The first term is the leaky current, which causes the membrane potential to decay toward the resting potential at a rate controlled by the membrane time constant. V_{rev} denotes the reversal potential, g represents the synaptic strength of the conductance, and G is the input conductance. The second term is the integrated term, indicating how the input conductance drives the membrane potential upward.

In most numerical implementations of the model, the variation of $V_{rest} - V$ is relatively small. Empirical studies on excitatory neurons indicate a resting potential of $V_{rest} = -70\ mV$, a reversal potential of $V_{rev} = 0\ mV$, and a firing threshold of $V_{th} = -50\ mV$. These values constrain $V_{rev} - V$ to the range $[50, 70]$, resulting in only a 1.4-fold difference between its maximum and minimum values. Therefore, $V_{rev} - V$ can be approximated as a constant and combined with the synaptic conductance strength into a single constant α, simplifying the original equation to:

$$\frac{dV}{dt} = \frac{V_{rest} - V}{\tau} + \alpha G$$

Assuming the input conductance G is proportional to the input current I_{origin}, we define $G = \beta I_{origin}$. Substituting this relationship and combining it with the membrane time constant τ, we obtain:

$$\frac{dV}{dt} = \frac{V_{rest} - V + \alpha\beta\tau I_{origin}}{\tau}$$

Letting $\alpha\beta\tau = scale_1$, we define a scaled input current $I = scale_1 I_{origin}$, corresponding to our scaling operation. Assuming a discretized time step $dt = 1$, we arrive at a simplified differential form of the LIF model:

$$\frac{dV}{dt} = \frac{I - (V - V_{rest})}{\tau}$$

After discretizations, considering the spiking behavior and introducing a buffer variable, the membrane potential update is given by:

$$U[t] = V[t] + \frac{1}{\tau}(I[t] - (V[t] - V_{rest}))$$

All other components of the model remain unchanged.

B Derivation of $scale_2$

Theorem 1 *Suppose $\mathbf{X}_{spike} \in \mathbb{R}^{T\times d_m}$ is a spike sequence with firing rate f, and $\mathbf{W}_Q$, $\mathbf{W}_K$ are matrices $\in \mathbb{R}^{d_m\times d}$ with zero mean and variances σ_1^2, σ_2^2, respectively. Assume that the elements in $\mathbf{X}_{spike}$, $\mathbf{W}_Q$, and $\mathbf{W}_K$ are mutually independent. Then, according to the computation of Spike-generated Multi-Head Attention, the mean of each element $(\mathbf{QK}^T)_{item}$ is zero, and we have the variance*

$$Var[(\mathbf{QK}^T)_{item}] = \frac{T-1}{T} dd_m^2 f^2 \sigma_1^2 \sigma_2^2 + \frac{1}{T}[f^2\sigma_1^2\sigma_2^2 dd_m(d_m - 1) + f\sigma_1^2\sigma_2^2 dd_m]$$

The second part can be neglected since T is relatively large in our setting (e.g., $T = 50$). Therefore, we approximate: $\text{Var}[(\mathbf{QK}^T)_{item}] \approx dd_m^2 f^2 \sigma_1^2 \sigma_2^2$. To ensure numerical stability, we normalize the result obtained from Spike-generated Multi-Head Attention by $scale_2 = \sqrt{\sigma_1^2\sigma_2^2 dd_m^2 f^2}$, yielding a variance of 1 for the output.

We now proceed to prove this result by introducing two supporting lemmas.

Lemma 1 *(Law of Total Variance): Suppose X and Y are measurable random variables on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then, the variance of X satisfies:*

$$Var[X] = Var_Y[\mathbb{E}[X|Y]] + \mathbb{E}_Y[Var[X|Y]]$$

Proof of Lemma 1: Notice that $\mathbb{E}_Y[\mathbb{E}[X|Y]] = \mathbb{E}[X]$, we can expand the variance as:

$$\begin{aligned}\text{Var}[X] =& \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ =& \mathbb{E}_Y[\mathbb{E}[X^2|Y]] - (\mathbb{E}_Y[\mathbb{E}[X|Y]])^2 \\ =& \mathbb{E}_Y[\text{Var}[X|Y]] + \text{Var}_Y[\mathbb{E}[X|Y]]\end{aligned}$$

Lemma 2 *Suppose a random variable X is sampled from a distribution with mean 0 and variance σ_0^2 with probability of p. Then, the variance of X is $p\sigma_0^2 + (1-p)\sigma_1^2$.*

Proof of Lemma 2: Let Y be a Bernoulli random variable indicating which distribution X is sampled from. Applying Lemma 1 (Law of Total Variance), we have

$$\text{Var}[X] = \text{Var}_Y[\mathbb{E}[X|Y]] + \mathbb{E}_Y[\text{Var}[X|Y]]$$

Since:

$$\begin{aligned}\text{Var}[X|Y=0] &= \sigma_0^2 \\ \text{Var}[X|Y=1] &= \sigma_1^2 \\ \mathbb{E}[X|Y=0] &= 0 \\ \mathbb{E}[X|Y=1] &= 0\end{aligned}$$

Combining them, we get:

$$\text{Var}[X] = 0 + p\sigma_0^2 + (1-p)\sigma_1^2$$

Returning to the original problem, we have the following expression:

$$\begin{aligned}&(\mathbf{Q}\mathbf{K}^T)_{ij} \\ =&\sum_{n=1}^{d}\sum_{m=1}^{d_m}\sum_{l=1}^{d_m}(\mathbf{X}_{spike})_{im}(\mathbf{W}_Q)_{mn}(\mathbf{X}_{spike})_{jl}(\mathbf{W}_K)_{ln}\end{aligned}$$

Due to independence and zero-mean assumptions, we immediately have:

$$\mathbb{E}[(\mathbf{Q}\mathbf{K}^T)_{ij}] = 0$$

The distribution of the spike sequence can be considered completely random when only the firing rate is known. Since $(\mathbf{X}_{spike})_{im}$ only takes values 0 or 1, it follows a Bernoulli distribution, and thus we have $\mathbb{P}((\mathbf{X}_{spike})_{im}=1) = f, \mathbb{P}((\mathbf{X}_{spike})_{im}=0) = 1-f$. Thus, its variance can be computed as $\text{Var}[(\mathbf{X}_{spike})_{im}] = f(1-f)$.

For simplicity, we omit the detailed derivation. The variance of $(\mathbf{Q}\mathbf{K}^T)_{ij}$ is as follows:

When $i \neq j$:

$$\text{Var}[(\mathbf{Q}\mathbf{K}^T)_{ij} = dd_m^2\sigma_1^2\sigma_2^2f^2$$

When $i = j$:

$$\text{Var}[(\mathbf{Q}\mathbf{K}^T)_{ij} = [d(d_m^2 - d_m)f^2 + dd_mf]\sigma_1^2\sigma_2^2$$

Applying Lemma 2 to compute the average variance across all elements in the attention matrix, we obtain:

$$\begin{aligned}\text{Var}[(\mathbf{Q}\mathbf{K}^T)_{item}] = &\frac{T-1}{T}dd_m^2f^2\sigma_1^2\sigma_2^2 + \\ &\frac{1}{T}[f^2\sigma_1^2\sigma_2^2dd_m(d_m-1) + f\sigma_1^2\sigma_2^2dd_m]\end{aligned}$$

C The metrics our experiments use

Assuming the true label of a sample is y_n, and the model's prediction is $\hat{y}_n$, our metrics are described as follows:

Acc-x is used to evaluate whether the model can divide the data into corresponding sentiment intervals, expressed as:

$$\texttt{Acc-x} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_n \in I(y_n)),$$

where N represents the number of samples involved in model evaluation, $\mathbb{1}(\cdot)$ is an indicator function. When the model prediction value $\hat{y}_n$ is within the sentiment interval of ground truth $I(y_n)$, it outputs 1, otherwise 0. `x` in `Acc-x` represents the number of sentiment classifications. As x increases, there are more sentiment intervals $\mathbb{1}()$, and the sentiment analysis becomes more detailed.

F1-score is used to evaluate the model's performance under data imbalance conditions. It combines precision P_c and recall R_c, which can be expressed as:

$$\begin{aligned}F1 &= \frac{2P_cR_c}{P_c+R_c} \\ P_c &= \frac{TP}{TP+FP} \\ R_c &= \frac{TP}{TP+FN}\end{aligned}$$

MAE is used to measure the accuracy of model regression and is expressed as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_n - \hat{y}_n|$$

Corr is used to measure the linear correlation between two variables. It is commonly applied in regression tasks to assess how well the predicted values align linearly with the ground truth. The formula is defined as:

$$\texttt{Corr} = \frac{\sum_{i=1}^{N}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i-\bar{\hat{y}})^2}},$$

where $\bar{y}$ and $\bar{\hat{y}}$ are means of y_i and $\hat{y}_i$. The closer the Corr value is to 1, the stronger the linear relationship between the predictions and the ground truth.

For all metrics except MAE, higher values indicate better performance.

Strength	Scale	Acc-5	Acc-3	Acc-2	F1	Corr	MAE	Spiking Rate
0	0.5	**55.03**	**72.15**	**80.27**	**80.36**	**0.7244**	**0.2976**	0.03
	1	52.80	70.70	78.92	79.03	0.6905	0.3156	0.07
	2	51.93	71.37	78.63	78.66	0.6917	0.3172	0.10
	3	51.93	71.76	77.85	77.90	0.6809	0.3212	0.11
0.01	0.5	**55.03**	**73.21**	**80.37**	**80.46**	**0.7247**	**0.3006**	0.03
	1	53.19	72.05	78.34	78.40	0.6925	0.3097	0.08
	2	53.19	71.95	78.34	78.37	0.6887	0.3170	0.10
	3	53.48	71.95	77.85	77.90	0.6865	0.3177	0.11
0.025	0.5	**54.93**	71.95	**79.98**	**80.10**	**0.7191**	**0.3085**	0.03
	1	52.61	**72.14**	79.21	79.26	0.6691	0.3156	0.08
	2	53.19	71.86	78.34	78.45	0.6884	0.3140	0.10
	3	51.74	71.37	78.34	78.42	0.6736	0.329	0.11
0.05	0.5	**54.06**	72.15	**80.56**	**80.57**	**0.7170**	**0.2992**	0.03
	1	53.68	**72.73**	78.92	78.67	0.7145	0.3032	0.09
	2	**54.06**	72.24	79.11	79.23	0.6984	0.3108	0.11
	3	52.71	70.79	77.76	77.85	0.6773	0.3207	0.11
0.1	0.5	28.82	41.49	57.16	50.49	0.1199	0.5051	0.03
	1	47.49	**71.37**	76.31	76.45	0.6540	0.3397	0.12
	2	**53.29**	71.08	**79.30**	**79.32**	**0.6548**	0.3288	0.16
	3	53.09	70.79	77.85	77.97	0.6663	**0.3229**	0.14

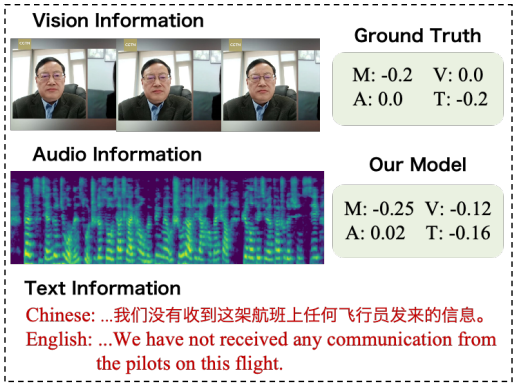
Table 4: The grid search results on SIMS Dataset. The bold and red bold values indicate the best results within each group and the overall best result in the table, respectively.

D Summary of Grid Search Results

We conducted a grid search experiment using input scaling factors $scale_2 = 0$ (blank control), 0.5, 1, 2, 3, and connection strengths of the connected LIF set to 0, 0.01, 0.025, 0.05, and 0.1. The final results are shown in Table 4.

E Details of Case Study

Figure 5 illustrates the detailed results of our method in two test samples from the SIMS dataset. Figure 5(a) shows a sample with natural sentiment type, and Figure 5(b) presents a difficult sample with inconsistent sentiment across modalities, where one modality is positive and the other is negative.



(a) Detail results in natural sample.

(b) Detail results in difficult sample.

Figure 5: Detail results of our method in two test samples of SIMS dataset.