Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Audiovisual Learning via Gated Spiking Cross-Attention

Yang Liu

sea@vip.henu.edu.cn

Henan University

Hang Wang

Henan University

Zhibing Li

Henan University

Jin Zhang

Henan University

Research Article

Keywords:

Posted Date: October 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10733322/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multimodal Audiovisual Learning via Gated Spiking Cross-Attention

Yang Liu[1], Hang Wang[1], Zhibing Li[1], Jin Zhang[1*]

[1*]College of Computer Science and Information Engineering, Henan University, Kaifeng, 475004, Henan, China.

*Corresponding author(s). E-mail(s): zhangjin@henu.edu.cn;
Contributing authors: sea@vip.henu.edu.cn; wanghang@henu.edu.cn;
2512454026@qq.com;

Abstract

Purpose Existing multimodal spiking Transformers lack a dedicated input-dependent mechanism for explicitly regulating information flow within cross-modal attention. Because the benefit of gating can depend on both its insertion point and driving signal, this study investigates explicit information-flow regulation, gating-position sensitivity, and driving-signal selection in multimodal Spiking Neural Networks (SNNs).

Methods We propose the Gated Spiking Cross-Attention Network (GSCANet), a unified audiovisual learning framework built on dual-stream Spikingformer-style encoding and bidirectional cross-modal interaction. Its core Gated Spiking Cross-Attention (GSCA) module introduces input-dependent regulation into the fusion process. Within this framework, we compare six gating placements spanning the Query, Key, and Value projections, attention output, feature concatenation, and block output, together with Q-only, V-only, and JointQV driving signals. The default configuration applies a head-specific multilayer perceptron (MLP) bottleneck gate to the attention output immediately before spike firing and drives it with the joint Query–Value representation.

Results Gating exhibits position sensitivity. Attention-output gating ($\boldsymbol{G_4}$) achieves 78.49% and 81.75% accuracy on CREMA-D and RAVDESS, improving over the No-Gate baseline by 2.15 and 1.05 percentage points, respectively. JointQV outperforms Q-only and V-only on all four datasets, indicating that effective modulation benefits from jointly representing interaction demand and transferred content.

Conclusion GSCANet provides a unified framework for explicitly regulated spiking audiovisual learning. Experiments within this framework show that

GSCA is position- and signal-dependent, supporting attention-output regulation before spike firing with JointQV as the driving signal.

Keywords: Spiking neural networks, multimodal audiovisual learning, information-flow regulation, gating placement, driving-signal selection, spiking Transformers

1 Introduction

Multimodal audiovisual learning exploits the complementarity of audio and visual information and has been widely studied for emotion recognition, event recognition, and natural human–computer interaction [2, 19, 33]. Cross-modal complementarity allows one modality to compensate for weaknesses in the other, such as visual occlusion or degraded auditory cues [23, 37]. Spiking Neural Networks (SNNs) [20], with their event-driven computation and sparse activation, provide a promising computational paradigm for energy-conscious multimodal perception [5, 10, 42]. Recent multimodal SNNs have progressed from low-level neuronal fusion [13, 17, 18] to attention-based global interaction [7, 10], enabling different modalities to exchange information in a shared spiking representation space.

Despite this progress, existing studies primarily emphasize encoder architectures, cross-modal topologies, and temporal modeling, while providing limited explicit control over how information is transferred inside cross-modal attention. In a typical spiking cross-attention block, Query, Key, and Value features are projected and aggregated before the resulting representation is passed to a spiking neuron. This interaction path contains no dedicated input-dependent mechanism for determining how strongly the current cross-modal representation should influence the subsequent spike response. This gap motivates a unified spiking audiovisual framework that supports cross-modal interaction while making information-flow regulation an explicit and testable component.

A direct approach to explicit information-flow regulation is to introduce multiplicative gating into the cross-attention pipeline [1, 11]. However, its effectiveness depends on two coupled design choices: where the gate is inserted and what information drives it. Candidate placements regulate different representations and determine whether modulation acts on continuous membrane-domain features or discretized spike responses. Meanwhile, a Query-only driver captures the interaction demand but not the supplied content, whereas a Value-only driver describes the transferred content without fully representing what the primary modality requires. Effective gating should therefore jointly consider its position relative to spike firing and a driving signal that represents both "what is needed" and "what is available."

Based on these considerations, we propose the Gated Spiking Cross-Attention Network (GSCANet), a unified framework that combines dual-stream Spikingformer-style encoding with stacked bidirectional cross-modal interaction. Its core GSCA module explicitly regulates cross-modal information flow through input-dependent

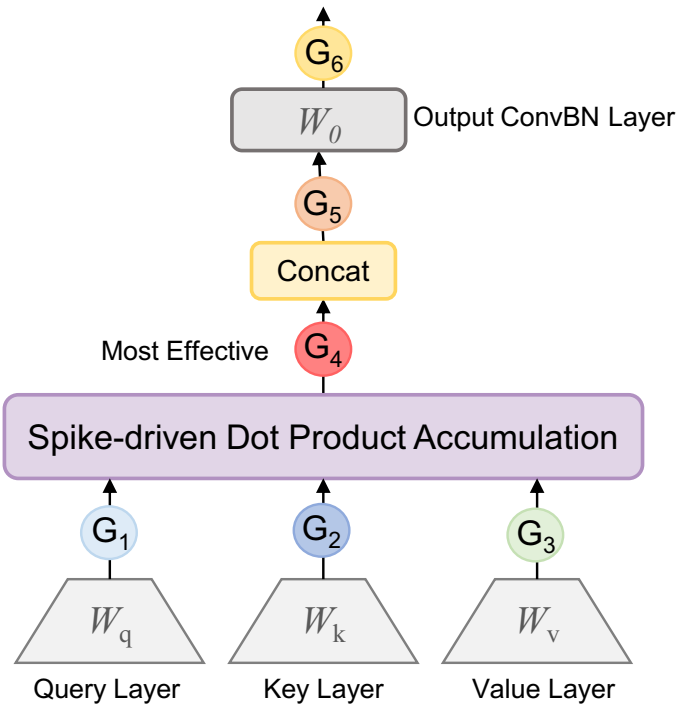


Fig. 1 Head-specific MLP bottleneck gating placements (G_1–G_6) in spiking cross-modal attention. The W_q, W_k, W_v, and W_o operators are kernel-1 ConvBN projections, and attention uses SDDPA without Softmax. Placements span Q/K/V projections (G_1–G_3), attention output (G_4), concatenation (G_5), and block output (G_6); G_4 is the default

multiplicative gating. Within this framework, we evaluate six head-specific multilayer perceptron (MLP) bottleneck placements spanning the Query, Key, and Value projections, attention output, feature concatenation, and block output (Fig. 1). The default G_4 configuration modulates the per-head attention output immediately before spike firing, preserving the completed cross-modal interaction while enabling continuous regulation before discretization. For gate generation, we compare Q-only, V-only, and JointQV signals; JointQV combines the interaction demand represented by the Query with the transferred content represented by the Value. Rather than modifying Query–Key matching, GSCA uses this signal to rescale the aggregated cross-modal representation before spike generation.

Experiments on the emotion benchmarks Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [2] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [19], as well as the general audio–visual benchmarks UrbanSound8K-AV and CIFAR10-AV [7, 26], reveal clear performance differences among gating placements. At the default time setting, G_4 improves over the No-Gate baseline by 2.15 and 1.05 percentage points on CREMA-D and RAVDESS, respectively, whereas some later placements do not yield comparable benefits. This result indicates that gating effectiveness depends on its position in the interaction

pipeline. The driving-signal ablation further shows that JointQV outperforms Q-only and V-only on all four datasets, supporting the use of both interaction demand and transferred content for modulation. Importantly, gating does not improve every placement or input condition, which further motivates the study of position sensitivity and driving-signal selection.

Our main contributions are summarized as follows:

- We propose GSCANet, a unified spiking audiovisual learning framework that integrates dual-stream Spikingformer-style encoding with stacked bidirectional cross-modal interaction.
- We design GSCA as the framework's core fusion module, introducing head-specific, input-dependent modulation to explicitly regulate cross-modal information flow before subsequent spike responses.
- Within the unified framework, we systematically evaluate six gating placements and three driving signals across four datasets. The results reveal position sensitivity, identify the pre-spike attention-output stage G_4 as an effective placement for the emotion tasks, and show the advantage of JointQV over single-source alternatives.

2 Related Work

2.1 Multimodal Audiovisual Classification with SNNs

Spiking Neural Networks (SNNs) exhibit strong spatio-temporal modeling potential in audio-visual classification (AVC) tasks. Existing literature broadly evolves across distinct paradigms based on the depth of cross-modal fusion. Early endeavors predominantly focused on low-level fusion at the neuronal and current levels, simulating biological information integration within neuronal dynamics or synaptic connections. For instance, spike-timing-dependent plasticity (STDP)-based methods [25, 39] explore unsupervised cross-modal associations, while CMCI [13] implements current-level integration. Recently, MISNet [18] introduced MLIF neurons to address audio-visual synchronization within a single neuron, establishing a hardware-friendly foundation for multimodal SNNs. Beyond audiovisual classification, spiking architectures have also been applied to multimodal emotion recognition, including NeuCube-based FusionSense [31], adversarially robust spiking Transformers (SPSNCVT) [3], and label-guided interactive networks (SMILE) [8].

Beyond these foundational low-level integrations, recent research has transitioned toward attention-based spiking architectures to capture global dependencies. Driven by the successful application of spiking self-attention (SSA) [42], frameworks such as the spiking multimodal Transformer (SMMT) [7], the event-based attention framework [17], and semantic-alignment and cross-modal residual learning (S-CMRL) [10] extend Transformer-style interaction to multimodal spiking representations. Within this progression, our work combines multi-frame visual input with a head-specific gate driven by concatenated Query and Value features. Unlike methods that modify attention scores, the default gate acts on the accumulated attention output immediately before spike firing.

Parallel to these structural enhancements, another prominent line of research addresses modal imbalance and temporal misalignment via temporal adaptive fusion. Han et al. [9] proposed balanced multimodal SNNs with online loss adjustment and convolutional time alignment, while Zhou et al. [41] introduced cross-modality attention for event–frame spatiotemporal fusion. A notable example is temporal attention-guided adaptive fusion (TAAF) [29], which dynamically scales weights along the temporal axis to stabilize multimodal learning. In artificial neural network (ANN)-based multimodal learning, related imbalance issues have been tackled by on-the-fly gradient modulation (OGM-GE) [23], adaptive gradient modulation (AGM) [15], modality-specific learning rates (MSLR) [36], and prototypical modal rebalancing (PMR) [6]. These methods are widely used as ANN baselines on CREMA-D (Table 1). While macro-level weight adaptation offers a valuable perspective, our work bridges the gap by exploring micro-level regulation within the attention mechanism itself. By embedding gating units within individual attention heads, we examine whether this fine-grained filtering can complement macro-adaptive strategies to provide a more stable spike integration environment for highly dynamic audio-visual streams.

2.2 Gating Mechanisms in Transformers

Gating mechanisms are widely used to regulate information flow and introduce multiplicative non-linearity in Transformer architectures. Historically, the utilization of gating began with feature filtering and non-linear modulation to substitute or augment traditional activation functions. For example, Gated Linear Units (GLU) [4] demonstrate that multiplicative interactions between twin linear paths yield superior modeling over single non-linear activations, while the Forgetting Transformer [16] selectively preserves temporal information by gating attention scores. Subsequent work systematizes GLU-style gating in Transformer feed-forward blocks, with SwiGLU and related variants [28] underpinning many large-scale language and vision–language backbones. These works motivate the use of multiplicative modulation, but they do not determine where an input-dependent gate should act in a softmax-free spiking cross-attention block.

As networks deepened, the research focus shifted from basic feature filtering toward positional and structural gating optimization within the attention pipeline. The gated attention unit (GAU) [11] shows that coupling gating with single-head attention rivals or outperforms multi-head attention. Similarly, recent studies on gated attention in large language models [24] reveal through systematic ablation that head-specific gating post-Scaled Dot-Product Attention (SDPA) alleviates training instability and enhances accuracy. Complementing these placement-centric studies, analyses of attention sinks show that normalization and aggregation choices can strongly shape softmax-attention behavior [34]. Because our spiking block omits Softmax, however, the appropriate gating stage must be validated separately. Meanwhile, data-dependent gates have been integrated into efficient attention alternatives such as gated linear attention [35], illustrating that gating is not limited to softmax SDPA pipelines. These explorations of gating positions and parameter granularities in ANNs motivate our systematic comparison at the projection, attention-output, concatenation, and block-output stages (G_1–G_6) within a spiking Transformer. Our comparison focuses on

placement-dependent accuracy and learned gate behavior rather than assuming that conclusions from softmax-based ANN attention transfer directly to spiking attention.

Beyond single-stream architectures, gating mechanisms have also been widely extended to cross-modal interaction and selection to coordinate multi-modal contributions. For instance, the gated multimodal unit (GMU) [1] utilizes internal neuronal gating units for dynamic inter-modal gain adjustment to mitigate modal conflicts, emphasizing gating as a "modal switch." Beyond static fusion units, recurrent multimodal designs employ gated recurrent unit (GRU)-style fusion to sequentially update a shared context, enabling explicit retain–overwrite control when new modalities arrive [30]. Inspired by this paradigm, our focus shifts from the mere presence of gating to the origin of the gating signal. Distinct from approaches driven by a single feature source, we test whether concatenating the Query (Q) and Value (V) provides a more informative basis for modulation by representing both interaction demand and transferred content.

3 Methods

3.1 Problem Definition and Preliminaries

We study multimodal audio-visual classification, including speech-emotion recognition (CREMA-D, RAVDESS) and general audio-visual categorization (UrbanSound8K-AV, CIFAR10-AV); the overall dual-stream architecture is shown in Fig. 2. Each sample contains paired audio and visual inputs: CREMA-D and RAVDESS provide synchronized recordings, whereas UrbanSound8K-AV and CIFAR10-AV use category-level audio–image pairings. Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v, y_i)\}_{i=1}^{n_t}$, where $\boldsymbol{x}_i^a \in \mathcal{X}^a$ and $\boldsymbol{x}_i^v \in \mathcal{X}^v$ denote the audio and visual inputs and $y_i \in \mathcal{Y}$ is the class label, the goal is to learn $f_{\boldsymbol{\theta}} \colon \mathcal{X}^a \times \mathcal{X}^v \to \mathcal{Y}$ with reliable cross-modal fusion:

$$\widehat{y} = f_{\boldsymbol{\theta}}\left(\boldsymbol{x}^a, \boldsymbol{x}^v\right). \tag{1}$$

To capture spatiotemporal dynamics in an event-driven manner, we build our method upon Spiking Neural Networks (SNNs) and adopt the Leaky Integrate-and-Fire (LIF) neuron as the basic computational unit. Following Spikingformer [40], the membrane potential dynamics and spike generation at discrete time step t are formulated as

$$H_t = V_{t-1} + \frac{1}{\tau_{\mathrm{m}}}\left(I_t - (V_{t-1} - V_{\mathrm{reset}})\right), \tag{2}$$

$$S_t = \Theta\left(H_t - V_{\mathrm{th}}\right), \qquad S_t \in \{0, 1\}, \tag{3}$$

$$V_t = H_t\left(1 - S_t\right) + V_{\mathrm{reset}} S_t, \tag{4}$$

where I_t is the synaptic input current, H_t and V_t denote the membrane potential before and after reset, respectively, τ_{m} is the membrane time constant, V_{th} is the firing threshold, and V_{reset} is the reset potential. $\Theta(\cdot)$ is the Heaviside step function. For stable end-to-end training, gradients through the reset operation are detached, and $\Theta(\cdot)$ is differentiated with a surrogate gradient.

3.2 Spikingformer-Style Spiking Patch Splitting and Unified Feature Representation

In SNNs, the temporal dimension T is fundamental to information encoding. To fully exploit the spatiotemporal dynamics of spike streams, we retain a Spiking Patch Splitting (SPS) encoder for the image-like visual frames and audio spectrograms, while updating its internal operations according to the pre-activation design of Spikingformer [40]. SPS maps heterogeneous inputs into a shared tensor space $\mathbb{R}^{T\times B\times N\times D}$ while ensuring that, except for the first encoding layer, each learnable projection receives binary spikes.

For video-based datasets (e.g., CREMA-D), each frame is encoded into spiking patches via SPS, and the video frame sequence is directly aligned with the SNN simulation time steps $t \in \{1,\dots,T\}$. This allows the model to capture the true temporal evolution and spatiotemporal continuity of physical events. For the audio modality, spectrograms are encoded into the same tensor format.

For datasets with static images or static spectrograms (e.g., UrbanSound8K-AV and CIFAR10-AV), single-frame features are replicated across all T time steps (Fig. 3). This design allows LIF neurons to accumulate membrane evidence from static inputs while maintaining the same temporal tensor interface used for dynamic tasks.

For modality $m \in \{a, v\}$, the j-th SPS stage is written as

$$\boldsymbol{F}_j^m = \mathrm{ConvBN}_j^m\left(\mathcal{P}_j\left(\mathrm{SN}_j^m\left(\boldsymbol{F}_{j-1}^m\right)\right)\right), \tag{5}$$

where every ConvBN_j^m comprises a 3×3 two-dimensional convolution with stride 1 and padding 1, followed by batch normalization. For a downsampling stage, $\mathcal{P}_j$ is a 3×3 max-pooling layer with stride 2 and padding 1; otherwise, it is the identity mapping. The first SPS layer directly encodes the non-spiking input with a 3×3 ConvBN layer and therefore omits the leading SN operation. Subsequent stages progressively increase the channel dimension to D. After the final stage, the feature map is flattened into N tokens and transposed to obtain $\boldsymbol{X}_a, \boldsymbol{X}_v \in \mathbb{R}^{T\times B\times N\times D}$. This SN–pooling–ConvBN ordering follows the pre-activation design of Spikingformer and differs from the ConvBN–SN ordering used in the previous backbone.

3.3 Gated Spiking Cross-Attention

Cross-modal interaction can benefit from an explicit mechanism that adjusts how strongly the aggregated representation contributes to the next spike response. We therefore introduce an input-dependent multiplicative gate and test whether its placement and driving signal affect information-flow regulation under a controlled architecture.

Given membrane-domain primary and auxiliary features $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{T\times B\times N\times D}$, GSCA first applies the Spikingformer pre-activation and then generates spiking query,

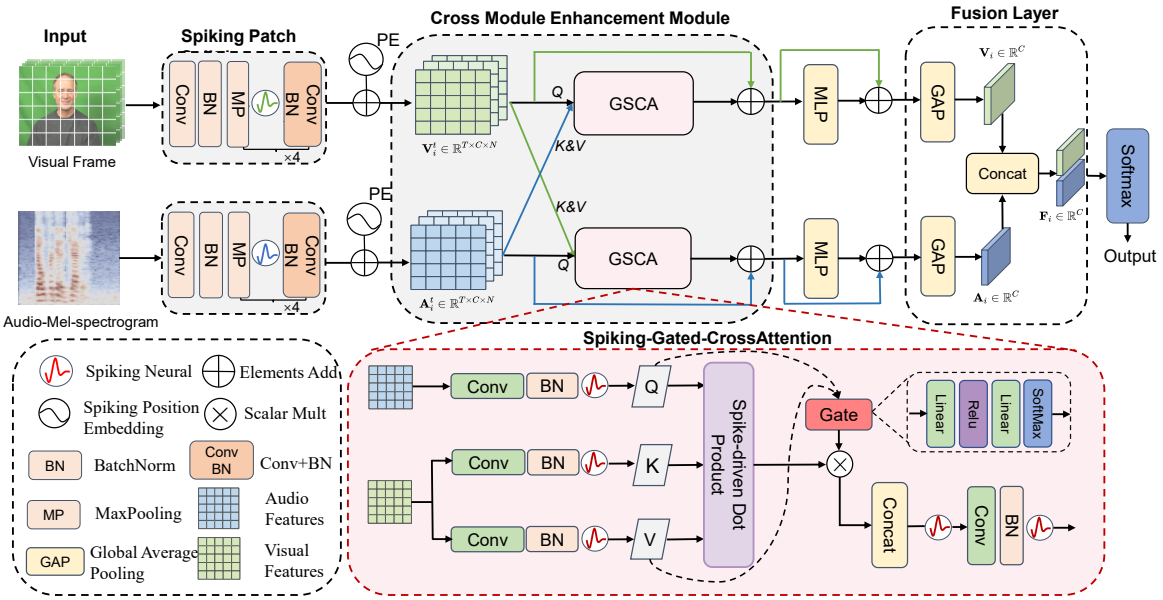


Fig. 2 Overview of GSCANet. Spikingformer-style Spiking Patch Splitting encoders process audio and visual inputs, followed by stacked bidirectional GSCA blocks with membrane-shortcut residual learning. The default G_4 gate uses JointQV to modulate attention outputs before spike firing; modality features are then pooled and concatenated for classification. Illustrative inputs are drawn from CREMA-D [2] and RAVDESS [19]

key, and value embeddings through pointwise ConvBN projections:

$$\begin{aligned} \overline{\boldsymbol{X}} &= \mathrm{SN}_x(\boldsymbol{X}), \qquad \overline{\boldsymbol{Y}} = \mathrm{SN}_y(\boldsymbol{Y}), \\ \boldsymbol{Q} &= \mathrm{SN}_q\big(\mathrm{ConvBN}_q(\overline{\boldsymbol{X}})\big), \\ \boldsymbol{K} &= \mathrm{SN}_k\big(\mathrm{ConvBN}_k(\overline{\boldsymbol{Y}})\big), \qquad \boldsymbol{V} = \mathrm{SN}_v\big(\mathrm{ConvBN}_v(\overline{\boldsymbol{Y}})\big), \end{aligned} \tag{6}$$

For these token projections, the feature layout is internally transposed to $\mathbb{R}^{T\times B\times D\times N}$, and each $\mathrm{ConvBN}_{q/k/v}$ is implemented by a one-dimensional convolution with kernel size 1 and stride 1, followed by one-dimensional batch normalization. Thus, the convolution performs a token-wise channel projection without changing the number of tokens. The features are then partitioned into H heads with $d = D/H$. Unlike conventional scaled dot-product attention, Spikingformer pre-activation spiking self-attention (PSSA) does not normalize $\boldsymbol{Q}\boldsymbol{K}^\top$ with Softmax. We refer to this operation as Spike-Driven Dot-Product Accumulation (SDDPA). The operation uses conventional matrix products algebraically, but its $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ operands are binary spike tensors emitted by the preceding spiking neurons; "spike-driven" therefore refers to the representation being accumulated rather than to a different multiplication rule.

For head h, SDDPA is formulated as

$$\boldsymbol{U}_h = s\left(\boldsymbol{Q}_h \boldsymbol{K}_h^\top\right)\boldsymbol{V}_h, \tag{7}$$

where s is a constant scaling coefficient that can be absorbed into the threshold of the following spiking neuron. For the default G_4 placement, the gate modulates this pre-spike accumulation:

$$\boldsymbol{G}_h = \sigma\left(g\left([\boldsymbol{Q}_h; \boldsymbol{V}_h]\right)\right), \quad \widetilde{\boldsymbol{U}}_h = \boldsymbol{U}_h \odot \boldsymbol{G}_h, \quad \boldsymbol{Z}_h = \mathrm{SN}_a\left(\widetilde{\boldsymbol{U}}_h\right), \tag{8}$$

where $[\cdot\,;\cdot]$ denotes feature concatenation, $\sigma(\cdot)$ is the Sigmoid function, and $\odot$ denotes element-wise multiplication. Thus, $\boldsymbol{G}_h$ changes the membrane-domain SDDPA output before SN_a emits a binary head output. The lightweight head-specific mapping is

$$g(\boldsymbol{\xi}) = \boldsymbol{W}_2\,\delta\left(\boldsymbol{W}_1\boldsymbol{\xi} + \boldsymbol{b}_1\right) + \boldsymbol{b}_2, \tag{9}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{(d/r)\times 2d}$, $\boldsymbol{W}_2 \in \mathbb{R}^{d\times(d/r)}$, $\delta(\cdot)$ is the rectified linear unit (ReLU), and r is the reduction ratio.

The binary outputs from all heads are concatenated and projected to the membrane domain by the output ConvBN layer:

$$\boldsymbol{Z}_{\mathrm{gsca}} = \mathrm{ConvBN}_o\left(\mathrm{Concat}\left(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_H\right)\right), \tag{10}$$

where ConvBN_o uses the same kernel-1, stride-1 one-dimensional projection and $\boldsymbol{Z}_{\mathrm{gsca}} \in \mathbb{R}^{T\times B\times N\times D}$. Following membrane-shortcut (MS) residual learning, no spiking activation is placed after this projection and before residual addition:

$$\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{Z}_{\mathrm{gsca}}. \tag{11}$$

The next GSCA or MLP projection receives $\mathrm{SN}(\boldsymbol{X}')$, so its learnable weights operate on binary spikes even though the shortcut state remains membrane-valued.

3.4 Systematic Exploration of Gating Configuration

To identify an effective integration path for gating within the cross-modal spiking Transformer, we conduct systematic ablations on gating placement topology and input driving sources under a unified formal framework.

3.4.1 Input Source Variants

To compare candidate driving signals, we define the gate input $\boldsymbol{I}_g$ as: (i) query-driven gating, $\boldsymbol{I}_g = \boldsymbol{Q}_h$, representing the primary modality's interaction demand; (ii) content-driven gating, $\boldsymbol{I}_g = \boldsymbol{V}_h$, representing the auxiliary content to be transferred; or (iii) joint-driven gating, $\boldsymbol{I}_g = [\boldsymbol{Q}_h; \boldsymbol{V}_h]$ (JointQV), which exposes both sources to the gate generator. The comparison tests whether this combined representation is more effective than either single-source alternative at the pre-firing modulation point.

3.4.2 Placement Variants

Given the selected driving source, we evaluate gating at four pipeline stages (Fig. 1): *projection* (G_1–G_3), *attention-output* (G_4), *concatenation* (G_5), and *block-output* (G_6). Unless noted otherwise, subsequent sections refer to placements by stage name and G_k index.

(1) Projection gating (G_1, G_2, G_3): filtering the pre-spike ConvBN responses and then regenerating binary Q/K/V embeddings:

$$\begin{aligned} G_1\colon&\quad \widetilde{\boldsymbol{Q}}_h = \mathrm{SN}_q\left(\boldsymbol{P}_h^q \odot \boldsymbol{G}_h\right), \\ G_2\colon&\quad \widetilde{\boldsymbol{K}}_h = \mathrm{SN}_k\left(\boldsymbol{P}_h^k \odot \boldsymbol{G}_h\right), \\ G_3\colon&\quad \widetilde{\boldsymbol{V}}_h = \mathrm{SN}_v\left(\boldsymbol{P}_h^v \odot \boldsymbol{G}_h\right), \end{aligned} \tag{12}$$

where $\boldsymbol{P}^q$, $\boldsymbol{P}^k$, and $\boldsymbol{P}^v$ are the corresponding ConvBN responses before their Q/K/V spiking neurons.

(2) Attention-output gating (G_4): modulating each head's membrane-domain SDDPA output $\boldsymbol{U}_h$ immediately before SN_a, as defined in Eq. (8).

(3) Concatenation gating (G_5): modulation after concatenating outputs $\boldsymbol{Z}_h$ from all attention heads and before the output ConvBN projection, focusing on cross-head channel selection:

$$\widetilde{\boldsymbol{Z}}_{\mathrm{cat}} = \mathrm{Concat}\left(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_H\right) \odot \sigma\left(g\left(\boldsymbol{I}_{\mathrm{global}}\right)\right), \tag{13}$$

where $\boldsymbol{I}_{\mathrm{global}} = \mathrm{MeanPool}_h([\boldsymbol{Q}_h; \boldsymbol{V}_h])$ pools JointQV features across heads before gating the concatenated SDDPA head outputs.

(4) Block-output gating (G_6): calibrating the projected GSCA residual branch immediately before MS residual addition:

$$\boldsymbol{X}'_{G_6} = \boldsymbol{X} + \boldsymbol{Z}_{\mathrm{gsca}} \odot \sigma\left(g\left(\boldsymbol{I}_g\right)\right). \tag{14}$$

Across the evaluated configurations, attention-output gating (G_4) with the JointQV source provides the most effective default for the two emotion benchmarks while remaining competitive on the two general audiovisual benchmarks.

3.5 Overall Framework and Loss Function

Our overall framework adopts a dual-branch Spikingformer backbone that encodes audio and video inputs into token sequences (Fig. 2). Each modality is first processed by the Spikingformer-style SPS encoder to obtain $\boldsymbol{X}^a, \boldsymbol{X}^v \in \mathbb{R}^{T\times B\times N\times D}$. Stacked audio-visual blocks perform bidirectional cross-modal interaction via GSCA (video queries audio and audio queries video in alternating blocks). For either modality, the l-th block follows the Spikingformer MS-residual order

$$\begin{aligned} \boldsymbol{X}'_l &= \boldsymbol{X}_{l-1} + \mathrm{GSCA}\left(\boldsymbol{X}_{l-1}, \boldsymbol{Y}_{l-1}\right), \\ \boldsymbol{X}_l &= \boldsymbol{X}'_l + \mathrm{ConvBN}_2\left(\mathrm{SN}_2\left(\mathrm{ConvBN}_1\left(\mathrm{SN}_1\left(\boldsymbol{X}'_l\right)\right)\right)\right). \end{aligned} \tag{15}$$

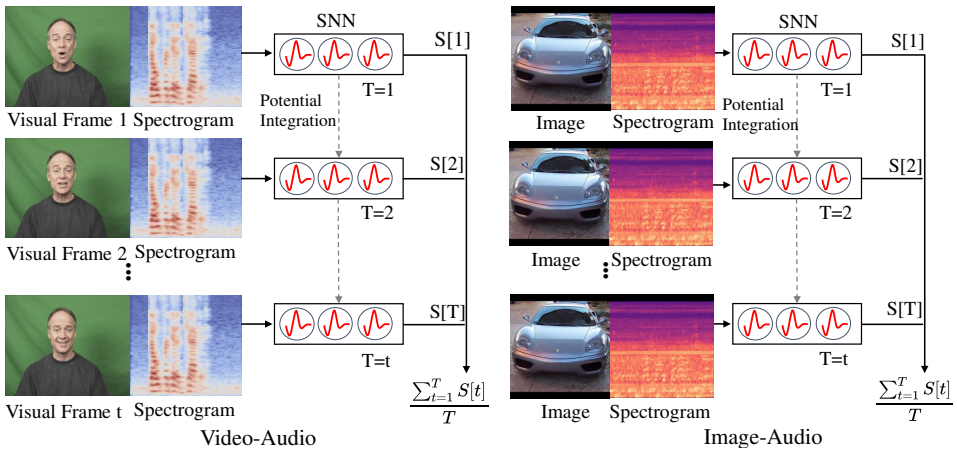


Fig. 3 Unified spiking encoding for heterogeneous audiovisual data. **Left:** video frames from CREMA-D [2] and RAVDESS [19] are aligned with simulation steps T. **Right:** static images from UrbanSound8K-AV and CIFAR10-AV [7] are replicated across T steps for temporal processing

Here, ConvBN_1 and ConvBN_2 form the Spikingformer MLP and are implemented as 1×1 two-dimensional convolutions with stride 1 on the reshaped spatial feature map (equivalently, token-wise channel projections). The first projection expands the channel dimension from D to $4D$, and the second restores it to D. The opposite branch is obtained by exchanging $\boldsymbol{X}$ and $\boldsymbol{Y}$. Spatial aggregation then reduces the token dimension, and temporal aggregation yields sample-level features for classification.

For optimization, all GSCANet variants are trained with the standard cross-entropy objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} = \mathrm{CE}(\boldsymbol{o}, y), \tag{16}$$

where $\boldsymbol{o}$ denotes the classification logits and y is the ground-truth class label.

4 Experiments

We first introduce the benchmark datasets and experimental configurations, then compare GSCANet with representative SNN and ANN methods. Experiments report clean-set accuracy, inference-time modality-failure behavior, and ablation results for each component.

4.1 Datasets

CREMA-D [2] is an audiovisual dataset for speech emotion recognition, containing 7442 video clips (2–3 seconds each) from 91 actors, covering six emotions: anger, happiness, sadness, neutrality, disgust, and fear. Following the preprocessing protocol in Guo et al. [7], we randomly partition the dataset with a 9:1 train/test ratio, yielding 6698 training samples and 744 test samples.

RAVDESS [19] (Ryerson Audio-Visual Database of Emotional Speech and Song) is a North American English corpus of acted emotional speech and song from 24

professional actors. Following common practice in multimodal emotion recognition, we use the speech full audio-visual subset (1440 synchronized clips, eight emotions) and randomly partition it with a 7:3 train/test ratio (1008 training, 432 test samples).

UrbanSound8K-AV [7] combines the UrbanSound8K audio corpus [26] with its corresponding image dataset, totaling 8732 audio-visual samples. Each sample consists of one high-resolution color image and a 4-second audio track. We follow Guo et al. [7] and randomly split the dataset into training and test sets with a 7:3 ratio.

CIFAR10-AV [7] pairs CIFAR-10 images with category-matched audio clips. Starting from 60,000 images and 10,000 audio segments, we obtain 10,000 audio-visual samples via one-to-one class pairing. Each sample contains one 32×32 color image and a 5-second audio signal across 10 object categories. We follow Guo et al. [7] and randomly split the dataset into training and test sets with a 7:3 ratio.

4.2 Experimental Settings

We follow the experimental protocol of prior work [7] and summarize it here for completeness. For the visual modality, static images are encoded by direct replication across SNN time steps. For CREMA-D, RAVDESS, UrbanSound8K-AV, and CIFAR10-AV, frames are resized to 128×128 pixels with random cropping for augmentation. Audio waveforms are normalized and resampled to 22,050 Hz, then transformed into 2D spectrograms via the short-time Fourier transform (STFT) with window length 512 and hop length 353. Amplitude spectrograms are log-scaled after adding a 10^{-7} offset and resized to match the image branch for dimensional consistency. CIFAR10-AV uses the same preprocessing and training configuration as UrbanSound8K-AV.

We follow the backbone design of Spikingformer [40] and adapt it to a dual-stream audio–visual architecture equipped with GSCA. All experiments are implemented using the Brain-Cog framework [38] on a single NVIDIA A100 graphics processing unit (GPU). We use the Adam optimizer [14] with an initial learning rate of 5×10^{-3}, training for 100 epochs with batch size 128. LIF neurons are initialized with membrane potential 0, firing threshold 1, and simulation time steps $T = 4$ for GSCANet unless noted otherwise. Surrogate gradients use a Sigmoid function $\sigma(x) = 1/\left(1 + \exp(-\alpha x)\right)$ with $\alpha = 4$ for end-to-end backpropagation through spiking layers. Unless stated otherwise, GSCANet, the No-Gate baseline, all gating ablations, and modality-failure tests use the same cross-entropy objective in Eq. (16), so comparisons among our variants isolate the effect of the GSCA design. Competing SNN methods in Table 1 retain their original training objectives. Following [7, 10], each configuration is trained once per split without multi-seed ensembling; margins below 1 percentage point on static benchmarks should be interpreted accordingly.

Use of generative AI. During preparation of this manuscript, the authors used OpenAI Codex to assist with structural and language revision and LaTeX editing. All AI-assisted output was reviewed and revised by the authors, who take full responsibility for the final manuscript.

4.3 Comparison with Representative Methods

Table 1 compares GSCANet with representative audio-visual SNNs and contextual ANN references. For brevity, we denote the cross-attention fusion method with deep metric learning proposed in [21] as AVCA-DML. We re-implement SNN baselines under the same preprocessing and dataset splits when public implementations are unavailable. An asterisk identifies a value taken from the cited paper; such values provide context but are not treated as protocol-matched results.

GSCANet obtains 78.49% on CREMA-D and 81.75% on RAVDESS. On CREMA-D, adding GSCA to the same backbone raises accuracy from 76.34% to 78.49%, giving a controlled gain of 2.15 percentage points over No-Gate. GSCANet also leads the listed SNN methods on both emotion benchmarks. Its CREMA-D margin over TAAF-SNNs is 0.94 points, although the two models use different time steps, and its RAVDESS margin over S-CMRL is 1.75 points. The higher AVCA-DML results follow a separate experimental protocol and are included only as external reference values.

On UrbanSound8K-AV and CIFAR10-AV, GSCANet reaches 98.47% and 99.00%, respectively, trailing MISNET-XL by only 0.49 and 0.26 points. Given the single-run setting, these sub-point gaps do not support a definitive ranking. The clearest benefit therefore appears on the naturally synchronized emotion datasets, where instance-level temporal correspondence may give the gate more meaningful cross-modal structure to regulate. The smaller margins on category-paired benchmarks are consistent with this interpretation, but do not establish it independently of dataset difficulty and composition.

4.4 Robustness under Modality Failure

We evaluate inference-time sensor loss by zero-masking either the audio or visual stream, following prior multimodal SNN studies [7, 10]. This diagnostic changes only the test input; the models are not retrained for missing modalities.

Table 2 shows an asymmetric response to sensor loss. When audio is masked, GSCA raises video-only accuracy by 2.69 percentage points on CREMA-D and 2.60 points on UrbanSound8K-AV. When vision is masked, the change is smaller and dataset-dependent: audio-only accuracy rises by 0.69 points on UrbanSound8K-AV but falls by 1.08 points on CREMA-D. GSCA therefore improves several failure conditions without providing symmetric or universal missing-modality robustness [10, 23].

RAVDESS provides a stricter counterexample. Under the same protocol, GSCA reduces video-only accuracy by 0.35 points and audio-only accuracy by 2.80 points, despite its strong clean-set result. A likely explanation is that the zeroed stream still enters cross-attention, moving the JointQV gate outside the input distribution learned with complete modality pairs. The gate should thus be interpreted as representation-level modulation rather than an explicit missing-modality recovery mechanism; its behavior depends on the dataset and on which sensor fails.

Table 1 Comparison of GSCANet with representative SNN and ANN methods on four datasets. Results marked with (*) are taken from reference papers; bold entries denote the best SNN accuracy per dataset under our evaluation protocols.

Dataset	Paradigm	Methods	Modality	Year	T	Accuracy (%)
CREMA-D	ANN	OGM-GE [23]	Multimodal	2022	–	62.20*
		MSLR [36]	Multimodal	2022	–	64.42*
		PMR [6]	Multimodal	2023	–	65.30*
		AGM [15]	Multimodal	2023	–	70.16*
		AVCA-DML [21]	Multimodal	2023	–	84.57*
	SNN	SMMT [7]	Multimodal	2023	4	66.53
		S-CMRL [10]	Multimodal	2025	4	73.25
		MISNET-XL [18]	Multimodal	2025	4	77.14
		TAAF-SNNs [29]	Multimodal	2025	3	77.55
		GSCANet	Multimodal	–	4	**78.49**
RAVDESS	ANN	Deep CNN [12]	Unimodal	2020	–	71.61*
		QCNN [22]	Unimodal	2021	–	77.87*
		MHCT [32]	Unimodal	2023	–	82.31*
		AVCA-DML [21]	Multimodal	2023	–	89.25*
	SNN	S-CMRL [10]	Multimodal	2025	4	80.00
		GSCANet	Multimodal	–	4	**81.75**
UrbanSound8K-AV	SNN	SMMT [7]	Multimodal	2023	4	96.85
		S-CMRL [10]	Multimodal	2025	4	98.13
		MISNET-XL [18]	Multimodal	2025	4	**98.96**
		GSCANet	Multimodal	–	4	98.47
CIFAR10-AV	SNN	SMMT [7]	Multimodal	2023	4	97.01
		S-CMRL [10]	Multimodal	2025	4	97.58
		MISNET-XL [18]	Multimodal	2025	4	**99.26**
		GSCANet	Multimodal	–	4	99.00

Table 2 Inference-time modality failure on CREMA-D and UrbanSound8K-AV. Audio zero and video zero denote the stream masked at test time.

Dataset	Method	Audio zero	Video zero
CREMA-D	No-Gate	62.63	47.98
	GSCANet	**65.32**	46.90
UrbanSound8K-AV	No-Gate	84.42	82.97
	GSCANet	**87.02**	**83.66**

4.5 Ablation Studies

4.5.1 Impact of Gating Placement

As shown in Table 3 and Fig. 4, gating placement clearly affects performance. Attention-output gating (G_4) achieves the best accuracy on CREMA-D and RAVDESS (2.15 and 1.05 percentage points above No-Gate, respectively) and remains competitive on UrbanSound8K-AV and CIFAR10-AV. Placement differences on the

Table 3 Ablation experiments on gating placement (G_1–G_6). Bold entries mark the best accuracy per dataset.

Method	CREMA-D	RAVDESS	UrbanSound8K-AV	CIFAR10-AV
No-Gate	76.34	80.70	98.66	96.76
Query (G_1)	77.15	80.35	98.62	98.37
Key (G_2)	78.09	77.89	98.70	98.77
Value (G_3)	77.28	80.35	98.62	98.37
Attn (G_4)	**78.49**	**81.75**	98.47	99.00
Concat (G_5)	77.95	80.70	98.67	98.26
Output (G_6)	75.26	76.14	**98.91**	**99.21**

latter two benchmarks are small (< 0.5 percentage points among the top configurations). Considering accuracy across all four datasets together with the placement-dependent gate behavior analyzed in Sec. 4.6, we adopt G_4 as the default GSCANet configuration. Although block-output gating (G_6) attains marginally higher accuracy on UrbanSound8K-AV and CIFAR10-AV (Table 3), the gap is below 0.5 percentage points, while G_6 collapses to a constant coefficient in the RAVDESS diagnostic and G_4 remains non-collapsed; we therefore use a single cross-benchmark default rather than dataset-specific placement tuning.

G_4 operates at the SDDPA output, after cross-modal dot-product accumulation and immediately before spike firing, enabling selective modulation of the completed cross-modal interaction rather than premature intervention.

An additional fixed-scale control clarifies the apparently favorable G_6 results on the two static benchmarks. On both UrbanSound8K-AV and CIFAR10-AV, the learned G_6 coefficients collapse to 0.5, and replacing the learned gate with a fixed coefficient of 0.5 reproduces the same accuracies (98.91% and 99.21%, respectively). Under Eq. (14), G_6 therefore reduces to $\boldsymbol{X}' = \boldsymbol{X} + 0.5\boldsymbol{Z}_{\text{gsca}}$ rather than performing input-dependent modulation. Its gains on these benchmarks should consequently be attributed to constant attenuation of the projected cross-modal residual branch, not to successful adaptive gating.

A plausible explanation is that the static, category-level audio–image pairings in UrbanSound8K-AV and CIFAR10-AV [7] can contain substantial redundant class evidence. Fixed attenuation may therefore improve the balance between the cross-modal residual and identity branches and reduce unnecessary downstream threshold crossings. By contrast, CREMA-D and RAVDESS contain synchronized, time-varying facial and vocal evidence [2, 19], for which uniform half-scaling can suppress useful instance-dependent interactions; this interpretation is consistent with the lower G_6 accuracies on both emotion benchmarks. Downstream spiking may also make the loss locally insensitive to changes in G_6 when the residual membrane state lies far from the firing threshold, weakening the gate gradient and leaving the Sigmoid output near $\sigma(0) = 0.5$. We treat this last point as a plausible optimization mechanism rather than a directly verified causal explanation.

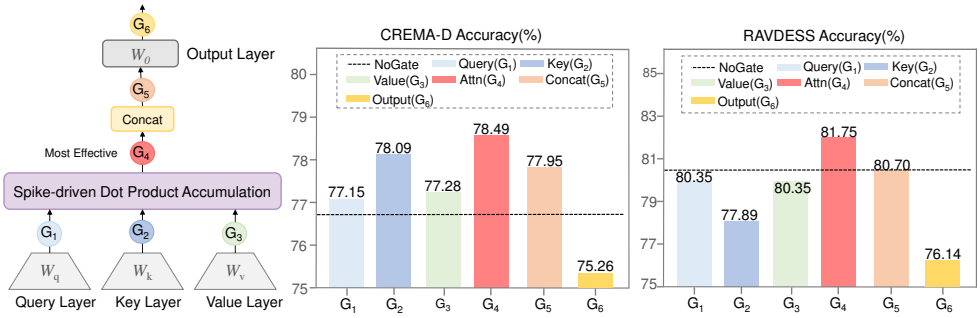


Fig. 4 Accuracy comparison of different gating placements on CREMA-D and RAVDESS. The compact labels Attn and Output denote attention-output (G_4) and block-output (G_6) gating, respectively. Attention-output gating achieves the best performance on these synchronized audiovisual emotion benchmarks

Table 4 Ablation experiments on gating input sources (Q-only, V-only, and JointQV).

Method	CREMA-D	RAVDESS	UrbanSound8K-AV	CIFAR10-AV
Q-only	75.13	78.94	98.15	97.68
V-only	76.74	78.94	98.27	97.41
JointQV	**78.49**	**81.75**	**98.47**	**99.00**

Table 5 Ablation experiments on single-frame versus multi-frame visual input on CREMA-D and RAVDESS.

Input Type	Methods	CREMA-D	RAVDESS
Single-frame	No-Gate	72.58	76.14
	GSCANet (G_4)	75.26	78.94
Multi-frame	No-Gate	76.34	80.70
	GSCANet (G_4)	**78.49**	**81.75**

4.5.2 Impact of Gating Input Sources

Table 4 compares Q-only, V-only, and JointQV drivers. JointQV is best on all four datasets; on CREMA-D it improves by 3.36 percentage points over Q-only. Under the evaluated settings, exposing both query intent and value content to the gate generator is more effective than using either source alone.

4.5.3 Impact of Temporal Resolution

Single-frame vs. multi-frame. As shown in Table 5 and Fig. 5, multi-frame input raises the No-Gate baseline from 72.58% to 76.34% (+3.76 percentage points) on CREMA-D under the evaluated split. GSCA yields gains in both settings (+2.68 percentage points for single-frame input and +2.15 for multi-frame input); jointly,

Table 6 Ablation experiments on spiking simulation time steps (T) on CREMA-D and RAVDESS.

Time Step	Methods	CREMA-D	RAVDESS
2	No-Gate	71.80	74.38
	GSCANet (G_4)	72.78	76.14
4	No-Gate	76.34	80.70
	GSCANet (G_4)	**78.49**	**81.75**
6	No-Gate	76.54	80.70
	GSCANet (G_4)	**78.98**	**82.80**

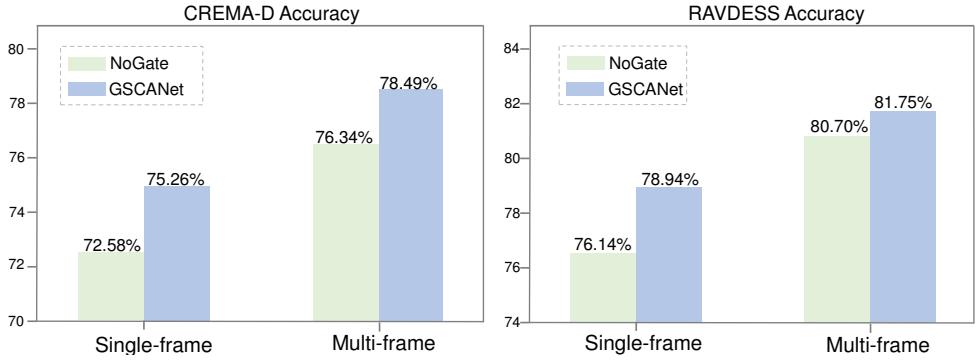


Fig. 5 Effect of single-frame vs. multi-frame visual input on CREMA-D and RAVDESS. Multi-frame input improves the baseline and synergizes with GSCA gating

multi-frame input and gating reach 78.49%. Thus, explicit modulation remains complementary to temporal input rather than replacing it.

Effect of temporal context length. With $T \in \{2, 4, 6\}$ aligned to input frames, Table 6 shows improving accuracy as context grows (e.g., CREMA-D No-Gate: 71.80% $\rightarrow$ 76.54%). GSCA remains effective at all evaluated lengths (+0.98, +2.15, and +2.44 percentage points on CREMA-D; +1.76, +1.05, and +2.10 on RAVDESS), with larger gaps at $T = 6$. These results show that the gate remains compatible with longer temporal context under the evaluated settings.

4.6 Diagnostic and Qualitative Analysis

We analyze the learned model along two complementary axes: (1) bidirectional descriptive statistics of the sigmoid gate coefficients across placements (Fig. 6), and (2) gradient-weighted class activation mapping (Grad-CAM) [27] visualization of discriminative regions (Fig. 7).

4.6.1 Placement-Dependent Gate Behavior

Fig. 6 summarizes the gate coefficients of all six placements (G_1–G_6) on the RAVDESS test set in both cross-modal directions. Panel (a) reports the directional means and

standard deviations, whereas panel (b) reports the absolute difference between the two directional means. The sigmoid outputs are bounded in $[0, 1]$ and are reported without renormalization.

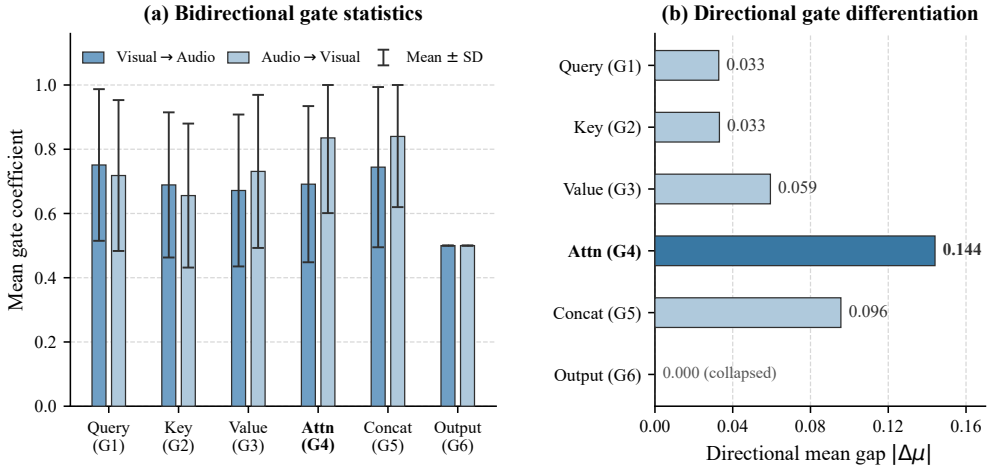


Fig. 6 Bidirectional gate statistics for G_1–G_6 on RAVDESS. (a) Mean sigmoid coefficients for visual-to-audio and audio-to-visual modulation; whiskers show one standard deviation. (b) Absolute directional mean gap $|\Delta\mu| = |\mu_{\mathrm{V}\to\mathrm{A}} - \mu_{\mathrm{A}\to\mathrm{V}}|$. G_4 shows the largest gap, whereas G_6 collapses to 0.5; these values are descriptive diagnostics

As shown in Fig. 6(a), G_1–G_5 retain nonzero coefficient dispersion, with standard deviations of approximately 0.22–0.25. The directional ordering also changes with placement: visual-to-audio means are slightly higher at G_1 and G_2, whereas audio-to-visual means are higher at G_3–G_5. Thus, gate placement affects both the operating level and directional response of the learned modulation.

At G_4, the mean coefficients are 0.691 for visual-to-audio modulation and 0.835 for audio-to-visual modulation, giving the largest directional mean gap (0.144) among the six placements, as highlighted in Fig. 6(b). The corresponding standard deviations remain nonzero (0.243 and 0.234). Because G_4 acts on the aggregated membrane-domain evidence before spike firing, this direction-sensitive behavior is consistent with head-specific regulation after cross-modal interaction. G_5 also retains nonzero dispersion and directional sensitivity, with a mean gap of 0.096. These statistics should therefore be interpreted as behavioral diagnostics rather than standalone evidence of superiority.

At G_6, both directions collapse to a constant coefficient of 0.5 (std. $= 0$), performing fixed half-scaling rather than adaptive modulation on RAVDESS. Its RAVDESS accuracy is 76.14%, below the 80.70% No-Gate baseline. Under the MS-residual formulation in Eq. (14), the block-output gate therefore degenerates into constant rescaling of the projected residual branch in this diagnostic.

Taken together, Fig. 6(a) and (b) show that gate placement changes the learned modulation behavior: G_4 combines non-collapsed coefficient variation with the strongest directional differentiation, whereas G_6 collapses on RAVDESS. Combined

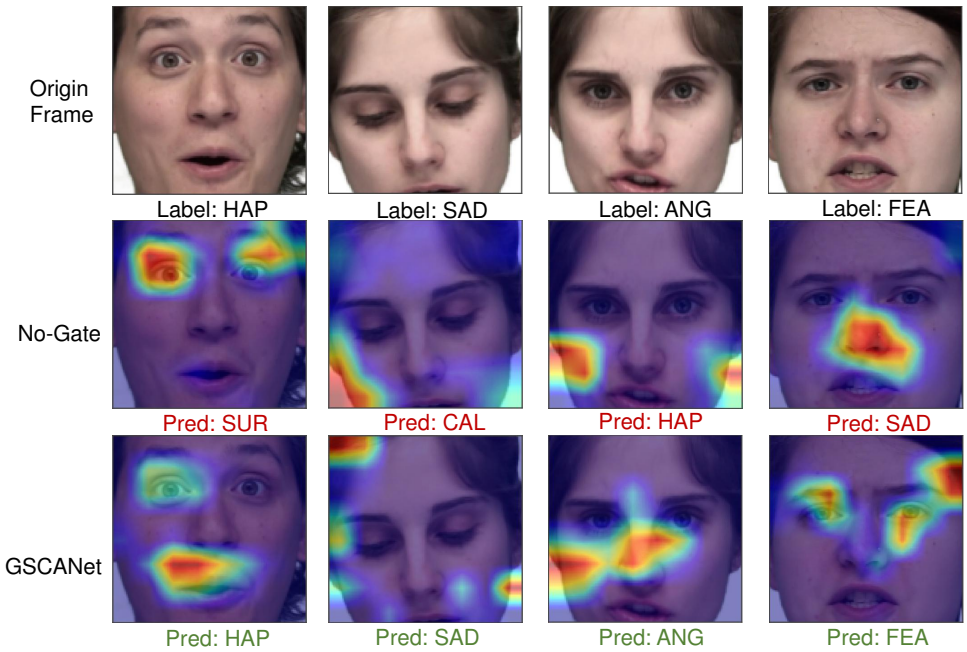


Fig. 7 Grad-CAM visualizations on selected RAVDESS samples [19] for four emotions: happiness (HAP), sadness (SAD), anger (ANG), and fear (FEA). Rows show the input frame with ground truth, No-Gate activations, and G_4-gated activations. On these examples, the G_4 variant produces more localized activation maps over facial regions

with the placement ablation in Table 3—rather than the coefficient statistics alone—these observations support G_4 as an effective cross-benchmark default, not as a universally optimal placement.

4.6.2 Grad-CAM Visualization

We apply Grad-CAM [27] to intermediate features in the GSCA-based cross-modal interaction block (not the final classifier layer) to highlight spatial regions associated with the predicted class via gradient-based localization; maps are therefore indicative rather than exact spatial-attention explanations. Fig. 7 compares the No-Gate baseline and GSCANet with G_4 on representative RAVDESS samples.

We select representative classes (HAP, SAD, ANG, FEA) where No-Gate fails but GSCANet with G_4 succeeds. In these selected examples, No-Gate activations are comparatively diffuse (e.g., HAP is misclassified as surprise (SUR), with a strong response around the eyes/eyebrows), whereas the G_4 variant produces more localized responses around the mouth or eye regions. Because G_4 modulates channels rather than explicit spatial attention, these maps are qualitative illustrations of changed feature use and should not be interpreted as a causal explanation of the accuracy gain.

5 Conclusion

This paper presents GSCANet, a unified framework for multimodal audiovisual learning with SNNs. Built on dual-stream Spikingformer-style encoding, it maps audio and visual inputs into compatible spiking representations and performs bidirectional interaction through GSCA. As the framework's core fusion module, GSCA introduces head-specific, input-dependent modulation into cross-modal attention, allowing aggregated information to be regulated before subsequent spike firing.

Within this framework, we systematically study six gating placements and three driving signals. The results reveal clear position sensitivity: attention-output gating at G_4 provides the strongest gated results on CREMA-D and RAVDESS under the default setting, while later placements may lose adaptive behavior. JointQV also outperforms Q-only and V-only across four datasets by combining interaction demand with transferred content. These findings demonstrate the feasibility of GSCANet for audiovisual classification and show that GSCA should be configured according to both the interaction stage and the information available to the gate.

Statements and Declarations

Funding. This work was supported by the Natural Science Foundation of Henan (No. 242300421218), the National Natural Science Foundation of China (No. 62176087), and the Scientific and Technological Innovation Team of Universities in Henan Province (No. 24IRTSTHN021).

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. Not applicable. This study is a secondary analysis of existing publicly available benchmark datasets and involved no new data collection, recruitment, or intervention involving human participants.

Consent to participate. Not applicable to the present study because no participants were recruited and no new human-participant data were collected by the authors.

Consent for publication. Figs. 2 and 7 contain facial images derived from CREMA-D and RAVDESS. The original RAVDESS publication reports that the depicted actors provided written informed consent for publication of their facial images [19]. Reuse of these images is subject to the applicable dataset licenses and publisher requirements.

Data availability statement. CREMA-D is publicly available from its official repository at https://github.com/CheyneyComputerScience/CREMA-D and is described in [2]. RAVDESS was accessed using the repository and download instructions provided at https://github.com/tuncayka/speech_emotion; the dataset is hosted on Zenodo at https://doi.org/10.5281/zenodo.1188976 and is described in [19]. UrbanSound8K-AV and CIFAR10-AV are publicly available through the SMMT repository at https://github.com/Guo-Lingyue/SMMT, with downloads provided at https://www.kaggle.com/datasets/lingyueguo/urbansound8k-av and https://www.kaggle.com/datasets/lingyueguo/cifar10-av, respectively. No new datasets were generated during this study.

Code availability. A public project repository has been created at https://github.com/lovecafe-132/GSSANet. The source code and Brain-Cog-based training scripts will be released in this repository.

Author contributions. Hang Wang: Writing–original draft, Software, Methodology, Data curation, and Conceptualization. Yang Liu: Funding acquisition, Writing–review and editing, Methodology, Conceptualization, and Resources. Zhibing Li: Software and Investigation. Jin Zhang: Data curation and Resources. All authors read and approved the final manuscript.

References

[1] Arevalo J, Solorio T, Montes-y Gómez M, et al (2017) Gated multimodal units for information fusion. Preprint at https://arxiv.org/abs/1702.01992

[2] Cao H, Cooper DG, Keutmann MK, et al (2014) CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing 5(4):377–390. https://doi.org/10.1109/TAFFC.2014.2336244

[3] Chen G, Qian Z, Zhang D, et al (2025) Enhancing robustness against adversarial attacks in multimodal emotion recognition with spiking transformers. IEEE Access https://doi.org/10.1109/ACCESS.2025.3544086

[4] Dauphin YN, Fan A, Auli M, et al (2017) Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning, PMLR, pp 933–941

[5] Deng S, Li Y, Zhang S, et al (2022) Temporal efficient training of spiking neural network via gradient re-weighting. In: International Conference on Learning Representations

[6] Fan Y, Xu W, Wang H, et al (2023) PMR: Prototypical modal rebalance for multimodal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 20029–20038, https://doi.org/10.1109/CVPR52729.2023.01918

[7] Guo L, Gao Z, Qu J, et al (2023) Transformer-based spiking neural networks for multimodal audiovisual classification. IEEE Transactions on Cognitive and Developmental Systems 16(3):1077–1086. https://doi.org/10.1109/TCDS.2023.3327081

[8] Guo M, Li W, Wang C, et al (2024) SMILE: Spiking multi-modal interactive label-guided enhancement network for emotion recognition. In: 2024 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6, https://doi.org/10.1109/ICME57554.2024.10688152

[9] Han J, Shen J, Xu Q, et al (2024) The balanced multi-modal spiking neural networks with online loss adjustment and time alignment. In: 2024 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6, https://doi.org/10.1109/ICME57554.2024.10687515

[10] He X, Zhao D, Dong Y, et al (2025) Enhancing audio-visual spiking neural networks through semantic-alignment and cross-modal residual learning. Preprint at https://arxiv.org/abs/2502.12488

[11] Hua W, Dai Z, Liu H, et al (2022) Transformer quality in linear time. In: International conference on machine learning, PMLR, pp 9099–9117

[12] Issa D, Demirci MF, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control 59:101894. https://doi.org/10.1016/j.bspc.2020.101894

[13] Jiang R, Han J, Xue Y, et al (2023) CMCI: A robust multimodal fusion method for spiking neural networks. In: International Conference on Neural Information Processing, Springer, pp 159–171, https://doi.org/10.1007/978-981-99-8067-3_12

[14] Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: International Conference on Learning Representations, URL https://arxiv.org/abs/1412.6980

[15] Li H, Li X, Hu P, et al (2023) Boosting multi-modal model performance with adaptive gradient modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, pp 22214–22224, https://doi.org/10.1109/ICCV51070.2023.02030

[16] Lin Z, Nikishin E, He XO, et al (2025) Forgetting transformer: Softmax attention with a forget gate. In: Proc. Int. Conf. Learn. Representations (ICLR)

[17] Liu Q, Xing D, Feng L, et al (2022) Event-based multimodal spiking neural network with attention mechanism. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 8922–8926, https://doi.org/10.1109/ICASSP43922.2022.9746865

[18] Liu X, Xia N, Zhou J, et al (2025) Towards energy-efficient audio-visual classification via multimodal interactive spiking neural network. ACM Transactions on Multimedia Computing, Communications and Applications 21(5):1–24. https://doi.org/10.1145/3721981

[19] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE 13(5):e0196391. https://doi.org/10.1371/journal.pone.0196391

[20] Maass W (1997) Networks of spiking neurons: The third generation of neural network models. Neural Networks 10(9):1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7

[21] Mocanu B, Tapu R, Zaharia T (2023) Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. Image and Vision Computing 133:104676. https://doi.org/10.1016/j.imavis.2023.104676

[22] Muppidi A, Radfar M (2021) Speech emotion recognition using quaternion convolutional neural networks. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6309–6313, https://doi.org/10.1109/ICASSP39728.2021.9414248

[23] Peng X, Wei Y, Deng A, et al (2022) Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 8238–8247, https://doi.org/10.1109/CVPR52688.2022.00806

[24] Qiu Z, Wang Z, Zheng B, et al (2025) Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)

[25] Rathi N, Roy K (2018) STDP based unsupervised multimodal learning with cross-modal processing in spiking neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence 5(1):143–153. https://doi.org/10.1109/TETCI.2018.2872014

[26] Salamon J, Jacoby C, Bello JP (2014) A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia, pp 1041–1044, https://doi.org/10.1145/2647868.2655045

[27] Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626, https://doi.org/10.1109/ICCV.2017.74

[28] Shazeer N (2020) GLU variants improve transformer. Preprint at https://arxiv.org/abs/2002.05202

[29] Shen J, Xie Y, Xu Q, et al (2025) Spiking neural networks with temporal attention-guided adaptive fusion for imbalanced multi-modal learning. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp 11042–11051, https://doi.org/10.1145/3746027.3755622

[30] Shihata Y (2025) Gated recursive fusion: A stateful approach to scalable multimodal transformers. Preprint at https://arxiv.org/abs/2507.02985

[31] Tan C, Ceballos G, Kasabov N, et al (2020) Fusionsense: Emotion classification using feature fusion of multimodal data and deep learning in a brain-inspired spiking neural network. Sensors 20(18):5328. https://doi.org/10.3390/s20185328

[32] Ullah R, Asif M, Shah WA, et al (2023) Speech emotion recognition using convolution neural networks and multi-head convolutional transformer. Sensors 23(13):6212. https://doi.org/10.3390/s23136212

[33] Wang D, Liu S, Wang Q, et al (2022) Cross-modal enhancement network for multimodal sentiment analysis. IEEE Transactions on Multimedia 25:4909–4921. https://doi.org/10.1109/TMM.2022.3183830

[34] Xiao G, Tian Y, Chen B, et al (2024) Efficient streaming language models with attention sinks. In: Proc. Int. Conf. Learn. Representations (ICLR)

[35] Yang S, Wang B, Shen Y, et al (2024) Gated linear attention transformers with hardware-efficient training. In: Proceedings of the 41st International Conference on Machine Learning, PMLR, pp 56501–56523

[36] Yao Y, Mihalcea R (2022) Modality-specific learning rates for effective multimodal additive late-fusion. In: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, pp 1824–1834, https://doi.org/10.18653/v1/2022.findings-acl.143

[37] Ye Q, Xu H, Ye J, et al (2024) mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13040–13051, https://doi.org/10.1109/CVPR52733.2024.01239

[38] Zeng Y, Zhao D, Zhao F, et al (2023) BrainCog: A spiking neural network based brain-inspired cognitive intelligence engine for brain-inspired AI and brain simulation. Patterns 4(8):100789. https://doi.org/10.1016/j.patter.2023.100789

[39] Zhang M, Luo X, Chen Y, et al (2020) An efficient threshold-driven aggregate-label learning algorithm for multimodal information processing. IEEE Journal of Selected Topics in Signal Processing 14(3):592–602. https://doi.org/10.1109/JSTSP.2020.2983547

[40] Zhou C, Yu L, Zhou Z, et al (2026) Spikingformer: A key foundation model for spiking neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 2236–2244, https://doi.org/10.1609/aaai.v40i3.37207

[41] Zhou S, Yang B, Yuan M, et al (2024) Enhancing SNN-based spatio-temporal learning: A benchmark dataset and cross-modality attention model. Neural Networks 180:106677. https://doi.org/10.1016/j.neunet.2024.106677

[42] Zhou Z, Zhu Y, He C, et al (2023) Spikformer: When spiking neural network meets transformer. In: International Conference on Learning Representations, URL https://iclr.cc/virtual/2023/poster/12139