SPIKEMOE: BRAIN-INSPIRED COMPETITIVE ROUTING FOR FLEXIBLE SPIKING MIXTURE-OF-EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spiking Neural Networks (SNNs) enable event-driven computation through biologically inspired dynamics at the neuronal scale, while Mixture-of-Experts (MoE) perform conditional computation through expert selection at the model scale. Integrating their strengths offers potential for flexible neural architectures. A key challenge, however, lies in designing an expert selection mechanism based on spiking activity. To address this, we introduce a spike-based k-WTA Router inspired by competition-inhibition observed in the hippocampal CA1 region. The router incorporates lateral inhibition and refractory period to select Top-K experts according to discrete spike counts. Building on this, we present SpikeMoE, a framework that integrates neuronal-scale spiking dynamics with model-scale expert selection. To address incomplete multisensory inputs in multimodal tasks, we further equip SpikeMoE with a two-stage missing-modality modeling module that combines empirical prototypes from an observed-modality pool with modality-specific learnable embeddings to construct missing-modality representations. Experiments on vision, language, and multimodal benchmarks demonstrate that SpikeMoE achieves state-of-the-art performance among the SNN baselines, matches or exceeds the performance of ANN counterparts, and maintains robustness across diverse missing-modality conditions. These results demonstrate a favorable trade-off between performance and energy efficiency, validating the integration of spiking dynamics with sparse expert computation and highlighting SpikeMoE as a promising approach to energy-efficient brain-inspired computing.

1 INTRODUCTION

The rapid advancement of artificial intelligence has given rise to increasingly complex neural network models, achieving success in numerous fields. However, modern neural networks consume vast amounts of computation and memory (Hwang et al., 2024). With the continuous growth in demand for intelligent systems, developing energy-efficient computing paradigms has become a critical priority. Nature provides an encouraging blueprint for high-efficiency computing. The human brain executes complex cognitive tasks with far greater efficiency than modern AI systems, all while consuming significantly less energy (Amunts et al., 2016). This remarkable efficiency stems from two key principles: at the microscopic level, biological neurons are activated by discrete, event-driven spikes rather than continuous signals (Roy et al., 2019), enabling sparse and asynchronous computation (Yao et al., 2023; 2024); at the macroscopic level, the brain exhibits distributed processing by selectively activating distinct regions for specific tasks (O'Doherty et al., 2021), enabling robust environmental perception through precisely coordinated multisensory integration.

Spiking Neural Networks (SNNs) (Izhikevich, 2003; Maass, 1997; Masquelier et al., 2008), considered the third generation of neural networks (Maass, 1997), simulate the dynamics of biological neurons at the microscopic level (Roy et al., 2019). The event-driven (Wei et al., 2024) characteristic of SNNs means that neurons fire spikes only when activated and remain silent otherwise (Akopyan et al., 2015). This mechanism increases the sparsity of synaptic computation and eliminates multiplication-accumulation (MAC) operations (Davies et al., 2018; Zhang et al., 2023; Wang et al., 2023b), thereby significantly improving energy efficiency. Furthermore, Mixture-of-Experts (MoE) (Mustafa et al., 2022) achieve dynamic expert selection through conditional computation (Shazeer et al., 2017; Jacobs et al., 1991; Chen et al., 1999; Yuksel et al., 2012), which aligns with the distributed and sparse processing at the macroscopic brain-region level (O'Doherty et al., 2021).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

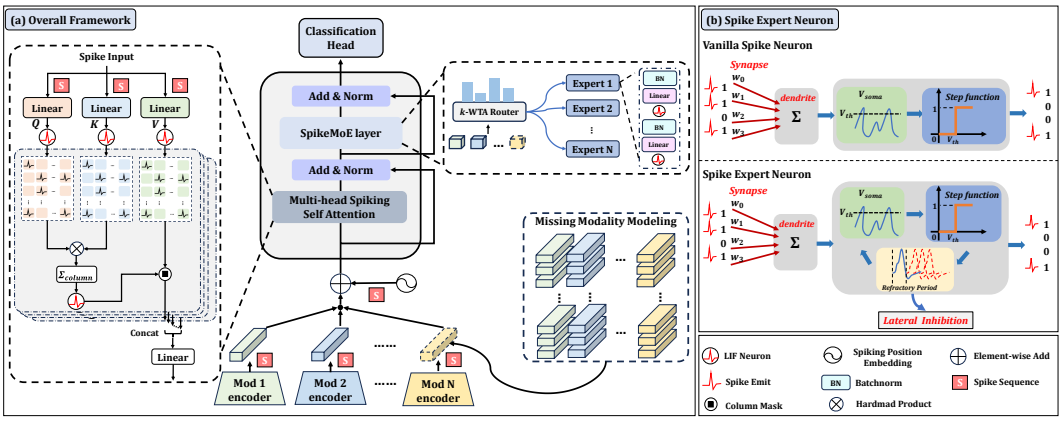


Figure 1: (a) Overall general framework; (b) Spike expert neuron.

Combining SNNs with MoE is promising because they provide complementary forms of efficient computation: SNNs exploit sparse, event-driven neuronal dynamics, whereas MoE performs conditional computation through expert selection. Nevertheless, developing a flexible spiking MoE presents two related challenges. At the mechanism level, conventional MoE routers typically generate continuous expert scores using softmax, whose exponentiation and global normalization operations are not naturally compatible with the discrete, event-driven computation of SNNs. A central challenge is therefore to perform effective expert selection directly based on spiking activity. At the perception level, multisensory inputs in real-world scenarios are frequently incomplete (Han et al., 2024; Glasser et al., 2016; Xu et al., 2023; Wu et al., 2024; Yuan et al., 2025; Zhang et al., 2022). Missing modalities deprive the router and experts of important contextual cues, potentially disrupting expert selection. Without appropriately representing the missing information and exploiting the available modalities, the resulting decisions can become fragile and suboptimal (Shaik et al., 2024; Cao et al., 2023; Tsai et al., 2019; Jiang et al., 2023; Peng et al., 2022).

To address these challenges, we propose SpikeMoE, a unified and flexible framework that integrates the strengths of SNNs and MoE. To resolve the internal bottleneck, we draw inspiration from hippocampal CA1 calcium imaging to propose k-WTA Router which is based on competition-inhibition mechanism (Topolnik & Tamboli, 2022). This approach achieves dynamic expert selection via sparse spike activation, effectively mitigating the issue of expert collapse. To overcome external perception challenges and improve robustness against incomplete inputs, we introduce a two-stage missing-modality modeling strategy. This enables the model to flexibly handle arbitrary combinations of modalities. The main contributions are summarized as follows:

- We introduce a k-WTA Router based on competition-inhibition mechanism. By incorporating lateral inhibition and refractory dynamics, the router performs Top-K expert selection according to spike counts and mitigates repeated within-input activation of a single expert.
- We develop a two-stage missing-modality modeling strategy that combines empirical information from an modality pool with modality-specific learnable embeddings, which enables the robust modeling of arbitrary modality combinations.
- By integrating the k-WTA Router with the missing-modality modaling strategy, we present SpikeMoE, a unified and flexible framework that connects neuronal-scale spiking dynamics with modal-scale sparse expert activation.
- Extensive evaluations show that SpikeMoE achieves performance comparable to or better than its ANN counterparts and establishes state-of-the-art results among the evaluated SNN baselines. Moreover, it remains robust under diverse missing-modality conditions while offering a favorable performance–energy efficiency trade-off.

2 Related Works

Multimodal learning in SNNs. With recent advances in neuromorphic computing, exploration of multimodal SNNs has gained momentum. We summarize main studies according to different

categories. **a. Audiovisual classification.** In (Liu et al., 2025), the introduction of MLIF neurons synchronizes and coordinates the spike activation of audiovisual signals within a single neuron; Liu et al. (Liu et al., 2022) proposed an event-based multimodal SNN comprising visual and auditory unimodal subnetworks and an attention-based cross-modal subnetwork for fusion. A spiking multimodal transformer was also proposed for multimodal audiovisual classification (Guo et al., 2023). **b. Emotion Recognition.** Tan et al. (Tan et al., 2020) employed an evolutionary architecture of multimodal SNNs based on the NeuCube framework; the Sliding Parallel Spiking Conv-ViT (Chen et al., 2025) integrated facial expressions, speech, and text, utilizing aligned heatmap features and multiscale wavelet transforms for precise feature extraction.

Multimodal learning in missing modality scenario. In many tasks, considering interactions among multimodal data can provide comprehensive understanding for downstream applications. MISTS (Zhang et al., 2022) employs a multi-time attention mechanism to encode irregular time series and utilizes multimodal cross-attention to fuse different modalities. In real-world scenarios, missing modalities pose a critical challenge. (Du et al., 2018; Shang et al., 2017) introduce generative adversarial networks to estimate missing modalities through semi-supervised learning; ShaSpec (Wang et al., 2023a) employs multiple loss functions to learn the proportion and specific representations of different modalities. However, the reconstruction networks in ShaSpec require multiple additional auxiliary loss functions, consuming substantial computational resources during optimization. Furthermore, recent studies like FuseMoE (Han et al., 2024) respect modal absence and aim to assign each combination of multimodal data to distinct experts.

3 PRELIMINARIES

3.1 SPIKING NEURON MODEL

SNNs achieve spike-driven processing using nonlinear spiking neurons. The Leaky Integrate-and-Fire (LIF) model is a widely adopted mathematical abstract for spike-driven behavior, achieving a balance between biological plausibility and computational efficiency (Maass, 1997). The dynamic model of LIF neuron is described as:

$$H[t] = V[t-1] + \frac{1}{\tau}(X[t] - (V[t-1] - V_{reset})), \tag{1}$$

$$S[t] = \Theta(H[t] - V_{th}), \tag{2}$$

$$V[t] = H[t](1 - S[t]) + V_{reset}S[t], \tag{3}$$

where τ is the membrane time constant, and $X[t]$ is the input current at time step t. When the membrane potential $H[t]$ exceeds the firing threshold V_{th}, the spike neuron will trigger a spike $S[t]$. $\Theta(\cdot)$ is the Heaviside step function which equals 1 for $v \geq 0$ and 0 otherwise. $V[t]$ represents the membrane potential after the trigger event which equals $H[t]$ if no spike is generated, and otherwise equals to the reset potential V_{reset}.

3.2 MIXTURE-OF-EXPERTS ARCHITECTURE

The Mixture-of-Experts (MoE) (Shazeer et al., 2017; Jacobs et al., 1991; Chen et al., 1999; Yuksel et al., 2012) enables selective expert activation, where only the top-k experts process each token. The MoE layer consists of a collection of m experts, $\mathbf{E} = \{\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_m\}$, and a router $\mathcal{R}_{\mathrm{A}}$ that selects the appropriate experts from this collection. Given the input sequence $\mathbf{X}_{\mathrm{A}}$, the output is formulated as the sum of the contributions of the top-k experts selected from the m candidates using the router:

$$\mathrm{y} = \sum_{k=1}^{K} \mathcal{R}_k(\mathbf{X}_{\mathrm{A}}) \cdot \mathbf{E}_k(\mathbf{X}_{\mathrm{A}}), \tag{4}$$

$$\mathcal{R}(\mathbf{x}) = \mathrm{TopK}(\mathrm{softmax}(\mathbf{X}_{\mathrm{A}}\mathbf{W}_{\mathcal{R}}, K)), \tag{5}$$

$$\mathrm{TopK}(\mathbf{v}, K) = \begin{cases} \mathbf{v} & \text{if } v \text{ is in the top } K \text{ elements} \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where $\mathbf{W}_{\mathcal{R}}$ denotes the weight matrix of the router. TopK combined with softmax$(\cdot)$ sets all elements of the routing vector to 0 except for the top K maximum values.

4 METHODS

4.1 K-WTA ROUTER IN SPIKEMOE

Building upon the original Transformer architecture, MoE introduces a substantial number of parameters and achieves conditional computation by calculating the routing probability via softmax function. However, SNNs must avoid multiplication, floating-point routing experts and softmax involving exponentiation and division do not adhere to the computational principles of SNNs. Therefore, it is necessary to design SNN-friendly gated routing mechanisms.

Competition-inhibition mechanism. Neuroscientists have observed lateral inhibition in calcium imaging results from the hippocampal CA1 region (Figure 2(a)) (Topolnik & Tamboli, 2022). It plays a vital role in information encoding and processing by enhancing neurons' specific responses to stimuli and improving spatial resolution in information processing (Topolnik & Tamboli, 2022; Jeong & Singer, 2022). It serves as a crucial mechanism for selective learning. Inspired by this phenomenon, we propose a novel spiking gating network based on a competition-inhibition mechanism, named the k-WTA (Winner-Take-All) Router. Specifically, it's constructed based on an improved LIF neuron model (Figure 1(b)), whose dynamics are described by the following discrete-time equation:

$$V[t] = V[t-1] \cdot (1 - \frac{1}{\tau}) + I[t], \tag{7}$$

where $V[t] \in \mathbb{R}^{B \times E}$ denotes the membrane potential tensor at time step t, B is the batch size, E is the number of experts, τ is the time constant, and $I[t]$ is the input current. To promote competitive selection among experts, we introduce a learnable lateral inhibition mechanism:

$$V_{inhibit}[t] = V[t] - S[t-1] \cdot (L \odot M) \cdot \gamma, \tag{8}$$

where $\odot$ is the Hadamard product, $L \in \mathbb{R}^{E \times E}$ is a learnable inhibition weight matrix, randomly initialized (with a standard deviation of 0.1) and set to 0 along the diagonal to prevent self-inhibition. $M \in \mathbb{R}^{E \times E}$ is the inhibition mask matrix, and γ is the inhibition strength coefficient. This competition-inhibition mechanism ensures that when an expert neuron fires, it suppresses the activation probability of other expert neurons, thereby achieving competitive selection.

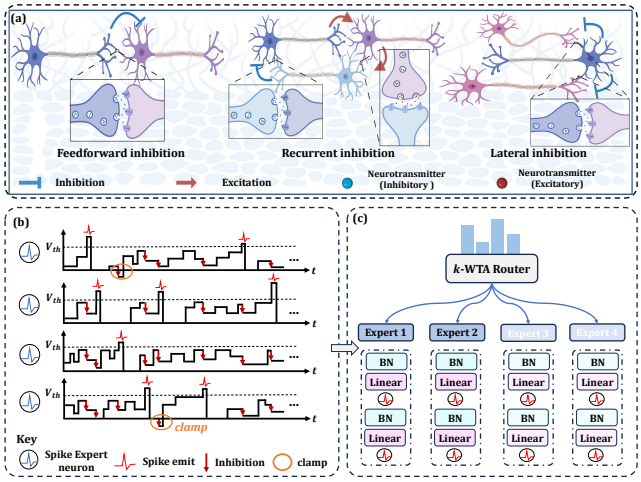


Figure 2: (a) Common inhibitory patterns in the hippocampus; (b) Competitive selection in k-WTA Router; (c) Expert selection.

Refractory period. During the competitive selection process, high-frequency neurons may undergo continuous triggering. This can cause the Top-K selection to degenerate into a monopoly by a single expert, leading to expert collapse. To avoid expert collapse, we introduce a refractory period. Under this mechanism, neuronal firing follows the following threshold-triggered mechanism:

$$S[t] = \begin{cases} 1 & \text{if } V[t] \geq V_{th}[t] \text{ and } R[t] \leq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where $S[t] \in \{0,1\}^{B \times E}$ denotes the spike firing state, while $R[t] \in \mathbb{R}^{B \times E}$ represents the refractory period counter. After a neuron fires a spike, its membrane potential resets and enters the refractory period (meanwhile, the other neurons continue to fire):

$$V[t] \leftarrow 0\,,\ R[t] \leftarrow t_{\text{refractory}} \text{ when } S[t] = 1. \tag{10}$$

The refractory period counter decrements over time, i.e., $R[t+1] = \max(0, R[t] - 1)$, ensuring neurons undergo a brief recovery period after firing a spike. To confine membrane potential and count

values within reasonable bounds, we employ clamp operations to non-negate membrane potential and count values. This prevents erroneous masking resulting from negative count values.

Expert selection and weight assignment. To achieve expert weight allocation based on spike count, we emulate the brain's multi-timescale characteristics (Yang et al., 2024) by introducing wta-timesteps t decoupled from the backbone's T. For an input tensor $X \in \mathbb{R}^{T\times B\times N\times C}$, T represents the extrinsic temporal dimension of the backbone model, characterizing the temporal structure of learned and encoded inputs. Meanwhile, t represents the simulation time step within the Router, used to compute the temporal count c_i via neurodynamics, which then informs the Top-K selection. Therefore, the spike count for each expert is:

$$c_i = \sum_{t=1}^{t} S_{b,i}[t], \text{ for } i = 1, 2, ..., E, \tag{11}$$

where i denotes the number of experts, $S_{b,i}$ represents the number of spike issued, and t is the simulation time step. Employing Top-K selection, the K experts with the highest spike counts are chosen: $\mathcal{E}_{\text{selected}} = \text{TopK}_{i\in[E]} c_i$. Expert weights are allocated proportionally based on spike counts:

$$w_i = \begin{cases} \frac{c_i+\varepsilon}{\sum_{j\in\mathcal{E}_{selected}}(c_j+\varepsilon)} & \text{if } i \in \mathcal{E}_{selected} \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

Here, ε represents a small random perturbation used to break ties when experts compete. The k-WTA Router's algorithm flow is shown in the Appendix A.6. The process of competitive selection by expert neurons is illustrated in the figure 2 (b) and (c). We demonstrate that the routing and expert-selection processes are spike-driven (proof in AppendixA.1.1), and by evaluating inference latency, we show that the introduced wta-timesteps does not create a computational bottleneck (Appendix A.5).

4.2 Two-stage Missing Modality Modeling Strategy

Previous work often employs filling or interpolating for the corresponding missing modality. However, this approach becomes cumbersome when considering interactions between modalities. We consider modality combinations under spike coding to ensure robust and flexible training as well as effective handling of missing modalities. Filling these modalities must be handled in such scenarios, rather than applying globally learned representation embeddings for each modality without considering the observed context.

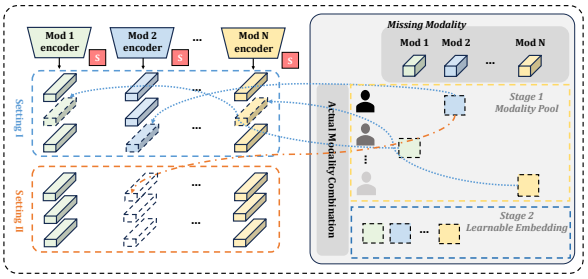


Figure 3: Two-stage missing modality modeling.

We propose a two-stage missing modality modeling strategy (Figure 3) that handles missing modalities through the adaptive fusion of a learnable modality pool and learnable embeddings. **Stage 1**: We define a binary mask $o_{i,n}$, where $o_{i,n}$ is defined as:

$$o_{i,n} = \begin{cases} 1 & \text{if modality is missing for sample} \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

where $o_{i,n} = 1$ indicates that the modality $M_i \in \mathbb{R}^d$ is missing, while $o_{i,n} = 0$ indicates that the modality is present. To ensure that at least one modality is retained for each sample, we enforce a constraint:

$$\forall i \in \{1, 2, ..., N\} : \sum_{n=1}^{M}(1 - o_{i,n}) \geq 1. \tag{14}$$

The modality pool is defined as $\mathcal{P}_k = \{M_{i,n} : o_{i,n} = 0,\ i \in \mathcal{D}_{\text{train}}\}$, containing all existing modalities in the training set where $o_{i,n} = 1$. When modality M_i is missing, we generate the imputed modality $\bar{M}_i = \frac{1}{K}\sum_{n=1}^{K} M_{i,n}$ by sampling K instances from the modality pool $M_{i,n} = [M_{i,0}, M_{i,1}, ..., M_{i,n}]$, where $M_{i,n}$ represents the i-th modality in the n-th instance of the pool.

Stage 2: When the modality pool cannot provide a suitable imputed modality (e.g., when the pool is empty, data quality is poor, or no matching contextual is available), a learnable embedding is used as the imputed modality. Given the number of modality combinations not fully observed, the total number of modalities in the modality pool is $\sum_{m=1}^{|M|-1}\binom{|M|}{m} = 2^{|M|} - 1$. Through the modality pool, the cascaded embeddings of all modalities for sample n, $h_n = [e_n^1, e_n^2, ..., e_n^i]$, are represented as follows:

$$e_n^i = \begin{cases} \text{Encoder}^i(n) & \text{if } o_{i,n} = 0 \\ M_{i,n} & \text{if } o_{i,n} = 1 \end{cases}, \tag{15}$$

where $e_n^i \in \mathbb{R}^d$ represents the embedding of the modality i from the sample n, with d being the hidden dimension. By doing so, the encoders for each modality can be trained without encountering unobserved, incomplete modalities. This approach simultaneously accounts for interactions between modalities and global representational information.

5 Experiments

5.1 Experiments Settings

Datasets. Given the generality of our approach, we evaluate it on the following datasets: ImageNet-1k (Deng et al., 2009), GLUE (Wang et al., 2018), CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), and UrbanSound8K-AV (Guo et al., 2023). Detailed informations are provided in Appendix A.2.1, while results on CMU-MOSI are available in Table 3 and 4. We conducted experiments on CMU-MOSEI and UrbanSound-AV, with results shown in Appendix A.3 and A.4.

Experimental Settings. On ImageNet-1K and GLUE, we evaluated its effectiveness by directly replacing the FFN layer in the baseline models with the k-WTA Router. To further assess SpikeMoE's robustness under missing modalities in real-world scenarios, we conducted extensive testing across three multimodal datasets. To simulate genuine modal deficiencies, we adopted the following two experimental settings: ***Setting I: Random Modality Dropout.*** A fixed percentage of modalities are randomly discarded for each instance, while ensuring at least one modality is retained for each instance in both the training and testing sets. ***Setting II: Asymmetric Modality Missingness.*** Half of the modalities are randomly discarded during training, while testing is performed only when each instance contains one or two modalities. For detailed configuration, see Appendix A.2

5.2 Primary Results

5.2.1 Effectiveness of k-WTA Router across Tasks

Results on ImageNet-1K. We replaced the FFN layer in the Spiking Transformer with a k-WTA Router and compared the improved model against the baseline. As shown in Table 1, the k-WTA Router consistently outperforms the baseline with slightly fewer model parameters. For instance, Spikformer-8-384 with the k-WTA Router outperformed the baseline by 2.71%, while Spike-driven Transformer-8-384 with the k-WTA Router achieved an improvement of 2.24%. Similar findings were observed on Spikingformer. We conducted a visual analysis of the conditional computation characteristics of the k-WTA Router (4 experts, Top-k=2). As shown in the Figure 4, for the same image, each expert assigns different computational regions to the routing process.

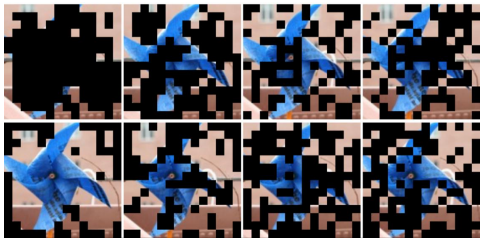


Figure 4: Visualization of routers as masks. The mask position (black) indicates router of 0 here. We show the dynamic sparsity for the selected experts (vertical) and time-steps (horizontal).

Results on GLUE Benchmark. For comparison purposes on the GLUE Benchmark, we replace the FFN layer in SpikeLM with k-WTA Router and compare the enhanced model against baselines. The ANN baselines include BERT (Devlin et al., 2019) and Q2BERT (with 2-bit weights and 8-bit activations) (Zhang et al., 2020), while the SNN baselines consist of SpikeBERT (Lv et al., 2023) and directly trained Spikingformer (Zhou et al., 2023). Following (Xing et al., 2024), we also implement

Table 1: Results on ImageNet. Model-L-D represents a model with L encoder blocks and D channels.

Methods	Architecture	Param (M)	Energy Consumption (mJ)	Time Step	Top-1 Acc (%)
Spikformer (Zhou et al., 2022)	Spikformer-8-384	16.81	12.43	4	70.24
	Spikformer-8-512	29.68	18.82	4	73.38
Spikformer+SEMM (Zhou et al., 2024)	Spikformer-8-384	16.05	11.50*	4	72.86
	Spikformer-8-512	28.22	17.63*	4	75.93
Spikformer+k-WTA Router	Spikformer-8-384	**16.03**	**11.39**	4	**73.22**
	Spikformer-8-512	**28.19**	**16.72**	4	**75.97**
Spike-driven Transformer (Yao et al., 2023)	SD-Transformer-8-384	16.81	3.90	4	72.28
	SD-Transformer-8-512	29.68	4.50	4	74.57
SD-Transformer+SEMM (Zhou et al., 2024)	SD-Transformer-8-384	16.05	3.75*	4	73.93
	SD-Transformer-8-512	28.22	4.23*	4	76.62
SD-Transformer+k-WTA Router	SD-Transformer-8-384	**16.03**	**3.54**	4	**74.52**
	SD-Transformer-8-512	**28.19**	**4.13**	4	**77.49**
Spikingformer (Zhou et al., 2023)	Spikingformer-8-384	16.81	4.69	4	72.45
	Spikingformer-8-512	29.68	7.46	4	74.79
Spikingformer+SEMM (Zhou et al., 2024)	Spikingformer-8-384	16.05	4.47*	4	73.58
	Spikingformer-8-512	28.22	6.94*	4	76.03
Spikingformer+k-WTA Router	Spikingformer-8-384	**16.03**	**4.27**	4	**74.37**
	Spikingformer-8-512	**28.19**	**6.63**	4	**76.92**

Note: * denotes our reproduced results.

spike-driven BERTs using PSN (Fang et al., 2023) and LIF (Gerstner et al., 2014) neuron. As shown in Table 2, the k-WTA Router improves SpikeLM performance by 1.2%. Furthermore, it narrows the performance gap with $\text{BERT}_{\text{base}}$ to 5.5%, significantly improving upon the original gaps of 6.7% for SpikeLM and 28.3% for LIF-BERT.

Table 2: Performance comparison on GLUE benchmark.

Model	Energy (mJ)	Time	$\text{MNLI}_{\text{-m/mm}}$	QQP_{F1}	QNLI	SST-2	CoLA	STS-B	MRPC_{F1}	RTE	Avg.
$\text{BERT}_{\text{base}}$ (Devlin et al., 2019)	51.41	/	83.8/83.4	90.5	90.7	92.3	60.0	89.4	89.8	69.3	83.2
BERT_{3L} (Devlin et al., 2019)	12.90	/	77.1/77.1	85.2	85.8	88.1	31.7	85.7	86.4	66.4	75.9
Q2BERT (Zhang et al., 2020)	/	/	47.2/47.3	67.0	61.3	80.6	0.0	4.7	81.2	52.7	49.1
ELMo (Peters et al., 2018)	/	/	68.6/-	86.2	71.1	91.5	44.1	70.4	76.6	53.4	70.2
SpikeBERT (Lv et al., 2023)	14.30	4	71.4/71.0	68.2	66.4	85.4	16.9	18.7	82.0	57.5	59.7
LIF-BERT* (Gerstner et al., 2014; Xing et al., 2024)	/	4	35.4/35.2	0.0	50.5	50.9	0.0	0.0	81.2	52.7	34.6
PSN-BERT* (Fang et al., 2023; Xing et al., 2024)	/	4	35.4/35.2	0.0	50.5	50.9	0.0	6.8	81.2	52.7	34.7
LIF-BERT (Gerstner et al., 2014; Xing et al., 2024)	7.98	4	56.8/55.2	70.0	60.6	80.6	14.6	20.0	82.3	53.8	54.9
Spikingformer (Zhou et al., 2023)	/	1	67.8/68.6	79.3	74.6	82.7	16.7	72.3	83.0	58.8	67.1
Spikingformer (Zhou et al., 2023)	/	4	70.2/70.6	80.9	79.5	83.9	12.8	77.0	83.0	62.1	68.9
SpikeLM (Xing et al., 2024)	13.74	4	77.1/77.2	83.9	85.3	87.0	38.8	84.9	85.7	69.0	76.5
SpikeLM+k-WTA Router	**12.45**	4	**77.2/77.4**	**85.6**	**86.6**	**88.2**	**42.2**	**86.4**	**86.2**	**69.7**	**77.7**

5.2.2 Robustness of SpikeMoE to Missing Modalities

Experiment Setting I: In Table 3 and Table 8, we implement Random Modality Dropout to study the sensitivity of model to missing modality. In this setting, we empirically observe that methods employing missing modal interpolation typically exhibit greater robustness under high levels of missing data, such as FuseMoE (Han et al., 2024) and our proposed approach. Although ShaSpec (Wang et al., 2023a) also estimates missing modes by sharing information, it is weak on the backbone model and difficult to optimize with multi-task objectives. On the other hand, compared to ANNs methods, our proposed approach demonstrates competitive performance across all metrics across most settings of CMU-MOSEI and CMU-MOSI; among SNNs methods, our approach achieves state-of-the-art performance. We present the UMAP projection of multimodal embeddings on the two datasets in Fig. 5 and 9, demonstrating that the impute modalities follow the same distribution as the original modalities. This validates the effectiveness of our method.



Figure 5: UMAP visualization of multimodal embedding space: CMU-MOSI.

Experiment Setting II: In Tables 4 and 9, we investigate which modality combinations are significant and how the resilience of our proposed method varies across different distributions under missing modalities. We find that in our proposed method, single modalities exhibit similar performance on F1 score and AUC during the inference stage, indicating the effectiveness of the missing modality filling strategy. Specifically, our proposed method also demonstrates robustness

Table 3: Experiment Setting I: Main results on CMU-MOSI. Top-3 results are highlighted as first, second and third.

Missing Ratio	Metric	ANNs				SNNs			Ours
		ShaSpec (Wang et al., 2023a)	TF (Zadeh et al., 2017)	mmFormer (Zhang et al., 2022)	FuseMoE (Han et al., 2024)	Weight Attention (Liu et al., 2022)	SCA (Guo et al., 2023)	S-CMRL (He et al., 2025)	SpikeMoE
0%	F1	41.78±0.64	43.28±0.36	50.22±0.20	57.02±0.39	50.41±0.44	46.95±0.21	49.01±0.29	59.59±0.23
	AUC	65.08±0.39	64.48±0.51	74.62±0.49	77.26±0.20	72.09±0.38	70.33±0.46	73.94±0.50	78.48±0.34
10%	F1	43.40±0.74	40.14±0.13	48.66±1.20	55.81±0.45	48.19±0.52	46.81±0.71	47.50±0.06	56.51±0.12
	AUC	65.45±0.56	61.54±0.66	73.21±0.76	77.05±0.93	74.68±0.17	70.51±0.02	72.12±0.27	77.45±0.67
20%	F1	39.15±0.13	42.64±0.30	48.51±1.15	53.42±0.58	47.46±0.02	47.43±0.29	47.35±0.35	53.73±0.92
	AUC	60.87±0.29	65.52±0.98	74.16±0.71	73.05±0.36	73.54±0.81	68.22±0.34	72.31±0.74	74.02±0.44
30%	F1	39.54±0.17	40.42±0.75	46.66±1.24	51.78±0.51	46.87±0.03	44.61±0.08	44.01±1.01	51.04±0.02
	AUC	60.80±0.44	60.57±0.99	71.25±0.57	72.38±0.56	72.31±0.11	68.48±1.03	70.21±0.05	73.15±0.35
40%	F1	35.32±0.32	38.69±0.42	42.22±0.44	48.78±0.42	48.45±0.20	41.81±0.37	42.11±0.64	48.99±0.51
	AUC	59.58±0.13	61.62±0.07	68.44±0.14	68.61±0.63	68.93±0.36	66.52±0.12	68.74±0.39	69.02±0.32
50%	F1	36.39±0.97	38.46±1.09	41.14±0.31	42.55±0.34	41.65±0.31	43.05±0.65	42.03±0.68	43.54±0.03
	AUC	55.60±0.12	60.49±0.80	61.87±0.67	67.28±0.28	65.81±0.05	64.49±0.43	65.71±0.43	67.30±0.71

Table 4: Experiment Setting II: Main results on CMU-MOSI. Top-3 results are highlighted as first, second and third.

Testing Modality			Metric	ANNs				SNNs			Ours
Video	Text	Audio		ShaSpec (Wang et al., 2023a)	TF (Zadeh et al., 2017)	mmFormer (Zhang et al., 2022)	FuseMoE (Han et al., 2024)	Weight Attention (Liu et al., 2022)	SCA (Guo et al., 2023)	S-CMRL (He et al., 2025)	SpikeMoE
✓	✓		F1	41.27±0.81	34.96±0.72	48.25±0.11	30.51±2.53	25.75±0.53	47.55±0.56	48.86±0.34	69.56±0.23
			AUC	61.28±0.88	59.37±1.12	72.72±0.08	55.35±1.11	54.39±0.12	71.98±0.43	73.22±0.06	74.29±0.18
✓		✓	F1	36.99±2.03	34.83±0.54	35.79±0.55	24.29±4.23	22.18±0.37	46.97±0.08	23.82±0.39	47.81±0.61
			AUC	55.79±2.53	57.87±0.32	52.21±0.31	56.24±1.96	50.45±0.08	68.56±0.23	51.55±0.28	69.04±0.18
	✓	✓	F1	43.72±1.64	37.25±0.53	45.53±0.33	20.68±2.14	49.40±0.23	19.18±0.47	48.09±0.21	49.47±0.38
			AUC	65.22±1.01	57.89±0.65	71.96±0.47	54.36±1.67	72.08±0.07	49.40±0.28	68.95±0.43	72.98±0.52
✓			F1	42.98±0.21	38.11±0.35	46.49±0.91	25.67±2.82	23.85±0.34	45.32±0.47	45.76±0.50	49.09±0.04
			AUC	55.10±1.59	58.87±0.44	71.01±0.53	55.94±1.51	52.49±0.38	70.21±0.55	70.32±0.03	72.29±0.09
	✓		F1	41.36±2.27	37.99±0.40	47.13±0.54	25.70±0.32	24.72±0.49	43.55±0.34	43.82±0.32	47.68±0.80
			AUC	61.60±4.80	60.28±1.42	73.77±0.44	55.45±0.95	53.87±0.33	71.02±0.27	71.52±0.04	72.30±0.26
		✓	F1	45.45±2.73	38.79±0.56	46.02±1.21	25.23±4.59	43.52±0.26	42.97±0.28	43.79±0.12	46.47±0.13
			AUC	66.85±3.01	57.19±0.72	68.04±0.94	56.04±1.91	54.08±0.63	67.06±0.02	68.22±0.15	69.03±0.08

to missing modalities across different combinations, meaning it maintains competitive performance regardless of which modality is absent. Furthermore, when evaluating the performance of all methods collectively, the video-audio modality combination yields poorer results, potentially due to weaker interactivity between these two modalities.

5.3 Ablation Study

Ablation on two-stage strategy. By progressively removing components designed in our proposed method, we investigated which parts of the model are actually useful (50% missing ratio on CMU-MOSI, as shown in Figure 7). We sequentially discarded the two stages of the missing modality modeling strategy and replaced missing modalities with zeros. When both stages were removed, F1 and AUC decreased by 5.12% and 5.27% on CMU-MOSI, and by 4.96% and 5.58% on CMU-MOSEI. Furthermore, we observed that removing stage 1 resulted in a greater decline than stage 2, suggesting that the modal interactions captured in the modality pool may play a more critical role in missing modalality modeling.

Table 5: Ablation on k-WTA Router.

Methods	Missing Ratio=50%	
	F1	AUC
Fully k-WTA Router	43.54 ± 0.03	67.30 ± 0.71
w/o Lateral Inhibition	41.37 ± 0.71	66.02 ± 0.49
w/o Refractory Period	42.50 ± 0.34	66.12 ± 0.56
w/o Inhibition Strength	40.41 ± 0.43	64.59 ± 0.80

Ablation on k-WTA Router. To validate the internal mechanisms of the proposed k-WTA Router, we conducted ablation studies at a 50% missing ratio on CMU-MOSI. As shown in Table 5, discarding key components—such as lateral inhibition or the refractory period—results in consistent performance degradation. Notably, eliminating the inhibition mechanism entirely (w/o Inhibition Strength) causes the most severe drop, with F1 and AUC falling by 3.13% and 2.71%, respectively. Furthermore, Fig.6 provides an intuitive explanation for this decline: without the refractory period, the routing process suffers from "expert collapse," where a single expert (expert 3) monopolizes the activation. These results collectively confirm that our brain-inspired competition-inhibition mechanisms are essential for preserving dynamic load balance and ensuring robust representation.

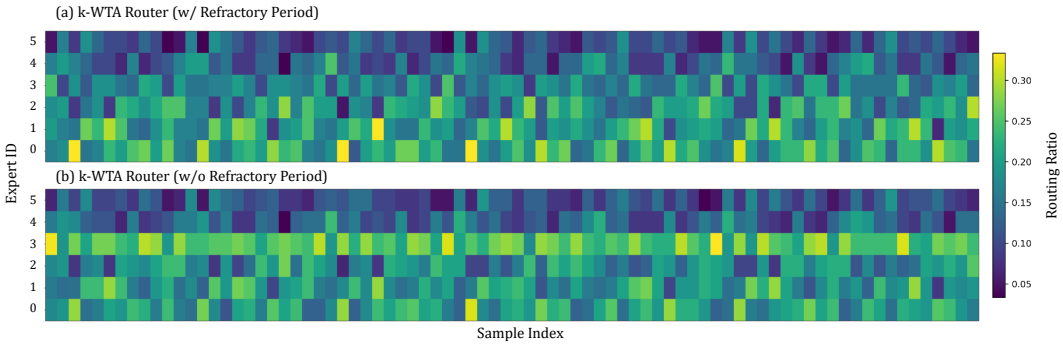


Figure 6: Visualization of expert activation.

5.4 Sensitivity Analysis

To demonstrate the robustness of SpikeMoE, we tested various settings of the number of experts, wta-timesteps, and inhibition strength under the configuration of missing ratio = 50% on CMU-MOSEI. We performed the sensitivity analysis by varying the values of the corresponding hyperparameters while keeping the other hyperparameters fixed at the values used in the experiments. Figure 8 presents the results of different parameters, indicating that the overall metrics remain relatively stable (dissused in A.2.7). In addition, we found that performance improved as the number and depth of experts increased, indicating that SpikeMoE demonstrates promising scalability.

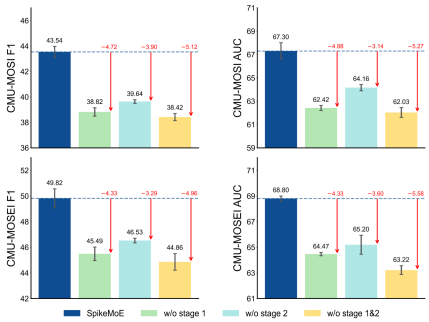


Figure 7: Ablation on two-stage strategy.

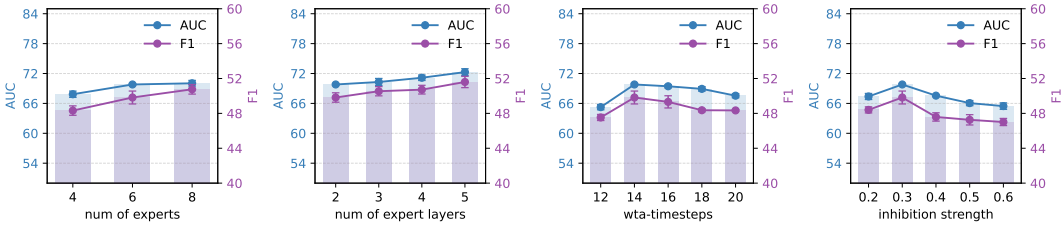


Figure 8: Sensitivity analysis on CMU-MOSEI (Missing ratio=50%).

5.5 Complexity Analysis

We compared the complexity of our proposed method and baseline models in terms of OPs (SOPs in SNN and FLOPs in ANN) parameters and energy consumption on CMU-MOSI (Missing ratio=50%). ShaSpec is the simplest model, yet its task performance is not comparable. In contrast, mmformer and S-CMRL are highly complex in terms of OPs and parameters, but their performance is not optimal. FuseMoE demonstrates performance comparable to our proposed method, yet it requires significantly more OPs and parameters. SpikeMoE is the optimal model that best balances complexity and performance. Detailed data are provided in Table 6.

6 Conclusion

In this work, we propose SpikeMoE, a unified framework that integrates Spiking Neural Networks (SNNs) with the Mixture-of-Experts (MoE) architecture. To address the inherent conflict between traditional softmax gating and SNN dynamics, we introduce a brain-inspired spiking gating network based on a competition-refractory-inhibition mechanism, which achieves dynamic expert selection through sparse spiking activation. Furthermore, to enhance the model's adaptability in complex scenarios, we propose a two-stage missing-modality modeling strategy. Extensive evaluations across vision, language, and multimodal datasets demonstrate SpikeMoE's robust performance. Future work will focus on applying SpikeMoE to a broader range of tasks, scaling the framework to larger architectures, and exploring its deployment on neuromorphic hardware.

AI USE STATEMENT

We used generative AI tools to improve the grammar, clarity, and wording of the manuscript, including assistance in drafting these statements. We did not use these tools to develop the research methodology, implement the model, generate or process experimental data, or analyze and interpret results. The authors reviewed and revised the AI-assisted text, verified its technical content against the research record, and take full responsibility for the final manuscript.

ETHICS STATEMENT

This study uses established image, language, and audiovisual benchmark datasets and involved no new recruitment of human participants or collection of personal data. The CMU-MOSI and CMU-MOSEI benchmarks contain recordings of people; their use follow the dataset providers' terms and applicable privacy requirements. Dataset composition and sentiment annotations may introduce biases that affect performance across populations.

REPRODUCIBILITY STATEMENT

We provide the model equations and implementation details needed to reproduce SpikeMoE in Section 4, including the k-WTA Router's membrane dynamics, lateral inhibition, refractory mechanism, spike-count-based expert selection, and two-stage missing-modality modeling strategy. Appendix A.6 presents pseudocode for the router, while Appendix A.1 provides the argument for its spike-driven characteristics. Section 5.1 defines the evaluation tasks and two missing-modality settings, including how modalities are removed during training and testing. Appendix A.2 documents the dataset partitions, evaluation metrics, modality-specific encoders, baseline methods, training objective, hardware, and energy-estimation procedure. Table 7 reports the principal hyperparameters and random seeds. We additionally report ablations and sensitivity analyses, with further results for CMU-MOSEI and UrbanSound8K-AV in Appendices A.3 and A.4. Together, these descriptions specify the experimental protocol and clarify that the reported energy consumption is a theoretical estimate rather than a hardware measurement.

REFERENCES

Filipp Akopyan, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam, Yutaka Nakamura, Pallab Datta, Gi-Joon Nam, et al. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. *IEEE transactions on computer-aided design of integrated circuits and systems*, 34(10):1537–1557, 2015.

Katrin Amunts, Christoph Ebell, Jeff Muller, Martin Telefont, Alois Knoll, and Thomas Lippert. The human brain project: creating a european research infrastructure to decode the human brain. *Neuron*, 92(3):574–581, 2016.

Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23555–23564, 2023.

Guoming Chen, Zhuoxian Qian, Dong Zhang, Shuang Qiu, and Ruqi Zhou. Enhancing robustness against adversarial attacks in multimodal emotion recognition with spiking transformers. *IEEE Access*, 2025.

Ke Chen, Lei Xu, and Huisheng Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural networks*, 12(9):1229–1252, 1999.

Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1):82–99, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, 2019.

Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 108–116, 2018.

Wei Fang, Zhaofei Yu, Zhaokun Zhou, Ding Chen, Yanqi Chen, Zhengyu Ma, Timothée Masquelier, and Yonghong Tian. Parallel spiking neurons with high efficiency and ability to learn long-term dependencies. *Advances in Neural Information Processing Systems*, 36:53674–53687, 2023.

Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Lingyue Guo, Zeyu Gao, Jinye Qu, Suiwu Zheng, Runhao Jiang, Yanfeng Lu, and Hong Qiao. Transformer-based spiking neural networks for multimodal audiovisual classification. *IEEE Transactions on Cognitive and Developmental Systems*, 16(3):1077–1086, 2023.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems*, 37: 67850–67900, 2024.

Xiang He, Dongcheng Zhao, Yiting Dong, Guobin Shen, Xin Yang, and Yi Zeng. Enhancing audio-visual spiking neural networks through semantic-alignment and cross-modal residual learning. *arXiv preprint arXiv:2502.12488*, 2025.

Sangwoo Hwang, Seunghyun Lee, Dahoon Park, Donghun Lee, and Jaeha Kung. Spikedattention: Training-free and fully spike-driven transformer-to-snn conversion with winner-oriented spike shift for softmax operation. *Advances in Neural Information Processing Systems*, 37:67422–67445, 2024.

Eugene M Izhikevich. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14 (6):1569–1572, 2003.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Nuri Jeong and Annabelle C Singer. Learning from inhibition: Functional roles of hippocampal ca1 inhibition in spatial learning and memory. *Current opinion in neurobiology*, 76:102604, 2022.

Runhao Jiang, Jianing Han, Yingying Xue, Ping Wang, and Huajin Tang. Cmci: A robust multimodal fusion method for spiking neural networks. In *International conference on neural information processing*, pp. 159–171. Springer, 2023.

Qianhui Liu, Dong Xing, Lang Feng, Huajin Tang, and Gang Pan. Event-based multimodal spiking neural network with attention mechanism. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8922–8926. IEEE, 2022.

Xu Liu, Na Xia, Jinxing Zhou, Zhangbin Li, and Dan Guo. Towards energy-efficient audio-visual classification via multimodal interactive spiking neural network. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(5):1–24, 2025.

Changze Lv, Tianlong Li, Jianhan Xu, Chenxi Gu, Zixuan Ling, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Spikebert: A language spikformer trained with two-stage knowledge distillation from bert. 2023.

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

Timothée Masquelier, Rudy Guyonneau, and Simon J Thorpe. Spike timing dependent plasticity finds the start of repeating patterns in continuous spike trains. *PloS one*, 3(1):e1377, 2008.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

John P O'Doherty, Sang Wan Lee, Reza Tadayonnejad, Jeff Cockburn, Kyo Iigaya, and Caroline J Charpentier. Why and how the brain weights contributions from a mixture of experts. *Neuroscience & Biobehavioral Reviews*, 123:14–23, 2021.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8238–8247, 2022.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202/.

Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, and Juan D Velásquez. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, 102:102040, 2024.

Chao Shang, Aaron Palmer, Jiangwen Sun, Ko-Shin Chen, Jin Lu, and Jinbo Bi. Vigan: Missing view imputation with generative adversarial networks. In *2017 IEEE International conference on big data (Big Data)*, pp. 766–775. IEEE, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Xinyu Shi, Zecheng Hao, and Zhaofei Yu. Spikingresformer: Bridging resnet and vision transformer in spiking neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5610–5619, 2024.

Clarence Tan, Gerardo Ceballos, Nikola Kasabov, and Narayan Puthanmadam Subramaniyam. Fusionsense: Emotion classification using feature fusion of multimodal data and deep learning in a brain-inspired spiking neural network. *Sensors*, 20(18):5328, 2020.

Lisa Topolnik and Suhel Tamboli. The role of inhibitory circuits in hippocampal memory processing. *Nature Reviews Neuroscience*, 23(8):476–492, 2022.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15878–15887, 2023a.

Yuchen Wang, Kexin Shi, Chengzhuo Lu, Yuguo Liu, Malu Zhang, and Hong Qu. Spatial-temporal self-attention for asynchronous spiking neural networks. In *IJCAI*, pp. 3085–3093, 2023b.

Wenjie Wei, Malu Zhang, Jilin Zhang, Ammar Belatreche, Jibin Wu, Zijing Xu, Xuerui Qiu, Hong Chen, Yang Yang, and Haizhou Li. Event-driven learning for spiking neural networks. *arXiv preprint arXiv:2403.00270*, 2024.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Xingrun Xing, Zheng Zhang, Ziyi Ni, Shitao Xiao, Yiming Ju, Siqi Fan, Yequan Wang, Jiajun Zhang, and Guoqi Li. Spikelm: Towards general spike-driven language modeling via elastic bi-spiking mechanisms. *arXiv preprint arXiv:2406.03287*, 2024.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

Wannan Yang, Chen Sun, Roman Huszár, Thomas Hainmueller, Kirill Kiselev, and György Buzsáki. Selection of experience for memory by hippocampal sharp wave ripples. *Science*, 383(6690): 1478–1483, 2024.

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

Man Yao, JiaKui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. *arXiv preprint arXiv:2404.03663*, 2024.

Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34, 2025.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

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

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

Jilin Zhang, Dexuan Huo, Jian Zhang, Chunqi Qian, Qi Liu, Liyang Pan, Zhihua Wang, Ning Qiao, Kea-Tiong Tang, and Hong Chen. 22.6 anp-i: A 28nm 1.5 pj/sop asynchronous spiking neural network processor enabling sub-o. 1 μj/sample on-chip learning for edge-ai applications. In *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 21–23. IEEE, 2023.

Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. Ternarybert: Distillation-aware ultra-low bit bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 509–521, 2020.

Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 107–117. Springer, 2022.

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

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

Zhaokun Zhou, Yijie Lu, Yanhao Jia, Kaiwei Che, Jun Niu, Liwei Huang, Xinyu Shi, Yuesheng Zhu, Guoqi Li, Zhaofei Yu, et al. Spiking transformer with experts mixture. *Advances in Neural Information Processing Systems*, 37:10036–10059, 2024.

A APPENDIX

A.1 MAIN PROPOSITIONS AND PROOFS

A.1.1 PROOF OF SPIKE-DRIVEN CHARACTERISTICS

We prove the spike-driven characteristics of the routing and expert-selection processes in k-WTA Router based on the definition proposed by (Shi et al., 2024).

Definition 1. *A spiking neural network is spike-driven if the input currents of all neurons satisfy the following form:*

$$I_i[t] = \sum_j w_{i,j} s_j[t] = \sum_{j, s_j[t] \neq 0} w_{i,j}, \tag{16}$$

where $I_i[t]$ is the input current of the i-th postsynaptic neuron at time step t, $s_j[t] \in \{0, 1\}$ is the spike output of the j-th presynaptic neuron, $w_{i,j}$ is the weight of the synaptic connection from neuron j to neuron i.

Proof. The membrane potential dynamics of the k-WTA Router are described by the discrete-time equation 7. Here, $V[t] \in \mathbb{R}^{B \times E}$ represents the membrane potential tensor, where B is the batch size and E is the number of experts. τ denotes the membrane potential time constant, and $I[t]$ is the input current tensor.

In the k-WTA Router, the input current $I[t]$ originates from the spike sequence of the backbone network. Let the output of the backbone network be the input tensor $X \in \mathbb{R}^{T \times B \times N \times C}$, where T is the backbone time step, B is the batch size, N is the sequence length or feature dimension, and C is the number of channels. In SpikeMoE, X is encoded as a binary spike sequence, with each element $s_j[t] \in \{0, 1\}$ representing the spike output of the presynaptic neuron, indexed by j.

For each neuron i in the k-WTA Router ($i = 1, 2, \ldots, E$), the input current $I_i[t]$ is computed as the weighted sum of the spike outputs from the presynaptic neurons:

$$I_i[t] = \sum_j w_{i,j} s_j[t], \tag{17}$$

where $w_{i,j}$ is the synaptic weight from presynaptic neuron j to k-WTA Router neuron i, satisfying Definition 1.

Moreover, the lateral inhibition and refractory period mechanisms introduced in the k-WTA Router do not affect the spike-driven nature of the input current:

- As shown in equation 8, the lateral inhibition mechanism acts on the membrane potential. Here, $S[t] \in \{0, 1\}^{B \times E}$ represents the spike output of the k-WTA Router neurons, $L \in \mathbb{R}^{E \times E}$ is the learnable inhibition weight matrix, $M \in \mathbb{R}^{E \times E}$ is the inhibition mask matrix, and γ is the inhibition strength coefficient. This mechanism adjusts the membrane potential but does not alter the form of the input current $I[t]$ calculation.
- As shown in equation 9, the refractory period mechanism affects membrane potential reset and spike emission. Here, $R[t] \in \mathbb{R}^{B \times E}$ is the refractory period counter. When a neuron spikes, the membrane potential is reset to $V[t] \leftarrow 0$ and $R[t] \leftarrow 2$, but this does not involve the calculation of the input current.

Therefore, the input current $I_i[t]$ is always composed of the weighted sum of the spike outputs from the presynaptic neurons, consistent with Definition 1. In summary, the routing and expert-selection processes in k-WTA Router exhibit spike-driven characteristics. □

A.2 EXPERIMENT DETAILS

A.2.1 DATASETS

ImageNet-1K. We use ImageNet-1K to validate the generalizability of the k-WTA Router on visual tasks. ImageNet (Deng et al., 2009) contains around 1,3 millon 1,000-class images for training and

50,000 images for validation. The input size of our experiments on ImageNet is set to the default 224×224.

GLUE. We applied the standard GLUE benchmark (Wang et al., 2018), using SpikeLM (Xing et al., 2024) as the baseline, to evaluate the performance of the k-WTA Router on natural language understanding tasks. We follow previous works and use the 8 subsets, including CoLA, STS-B, MRPC, RTE, QQP, MNLI and QNLI for classification or regression in different scenes. For evaluation, we follow BERT (Devlin et al., 2019) and report F1 scores for QQP and MRPC datasets; Spearman correlations for the STS-B dataset; and accuracy scores for other datasets.

CMU-MOSI. CMU-MOSI (Zadeh et al., 2016) is a multimodal sentiment analysis dataset comprising audio, video, and text modalities. It comprises 2,199 monologue video samples, with 1,284 used for training, 229 for validation, and 686 for testing. Acoustic and visual features are sampled at 12.5Hz and 15Hz.

CMU-MOSEI. CMU-MOSEI (Zadeh et al., 2018) utilizing the same modalities as CMU-MOSI. It includes 22,856 YouTube movie review clips, with 16,326 for training, 1,871 for validation, and 4,659 for testing. Acoustic and visual features are sampled at 20Hz and 15Hz. Sentiment labels on both datasets range from -3 (highly negative) to 3 (highly positive).

UrbanSound8k-AV. The UrbanSound8K-AV dataset (Guo et al., 2023) merges the UrbanSound8K audio dataset (Salamon et al., 2014) with its corresponding images, comprising 8732 audiovisual samples. Each sample consists of a high-resolution color image and a 4-second audio signal. We follow Guo et al. (Guo et al., 2023) by randomly dividing the dataset into training and test sets in a 7:3 ratio.

The image classification (ImageNet-1K) experiments were conducted on 4 NVIDIA H100 GPUs, while the remaining experiments were conducted on 8 NVIDIA RTX 4090 GPUs. To evaluate the generalizability of the k-WTA Router, we conducted experiments by replacing the FFN layer in the baseline model with the k-WTA Router. For assessing robustness under missing modalities, we conducted experiments using two experimental settings (refer to Sec.5.1) on CMU-MOSI and CMU-MOSEI, respectively, while only experiment setting I was employed for experiments on UrbanSound8k-AV. The comparison between CMU-MOSI and CMU-MOSEI includes all 8 baseline methods, while the comparison for UrbanSound8k-AV only includes the SNN baseline method. For ease of comparison, we use F1 scores and AUC values as evaluation metrics for each method on CMU-MOSI and CMU-MOSEI; we use accuracy as the evaluation metric for each method on UrbanSound8k-AV.

A.2.2 Model Details

In this section, we supplement the model details not covered in the main text.

Modality-Specific Encoders. To eliminate the confusing effects arising from different encoder architectures, we employ a modality-specific encoder with the same architecture across all models, as variations in encoder design significantly impact final performance.

Spiking visual encoder employs a dual-path hierarchical architecture, performing spike encoding separately on the channel dimension and feature dimension. Given visual data $X_v \in \mathbb{R}^{B\times C\times F}$, where B is the batch size, C is the number of channels, and F is the feature dimension, the encoding process can be represented as:

$$S_c = \mathcal{SN}(W_c\,\mathrm{LN}(X_v^\top)), \tag{18}$$

$$S_f = \mathcal{SN}(W_f\,\mathrm{LN}(S_c^\top)), \tag{19}$$

where $\mathcal{SN}$ denotes spiking neuron layer, $W_c \in \mathbb{R}^{N\times C}$ denotes the channel projection weights, $W_f \in \mathbb{R}^{N\times C}$ denotes the feature projection weights, and LN represents LayerNorm. The final output is $\mathcal{E}_v = \{S_f\}^T \in \mathbb{R}^{T\times B\times N\times C}$.

Spiking audio encoder employs temporal slicing and spiking convolution, enhancing local feature extraction capabilities by segmenting long-duration signals and introducing spatio-temporal convolution.

For the audio signal $X_a \in \mathbb{R}^{B\times S\times E}$, temporal segmentation is first performed:

$$P = \lceil S/L_P \rceil, \tag{20}$$

$$\tilde{X}_a = \text{Padding}(X_a, P \times L_P - S), \tag{21}$$

$$X_{patch} = \text{Unfold}(\tilde{X}_a, \text{kernel size}, \text{stride}). \tag{22}$$

Subsequently, temporal feature extraction and pulse coding are performed:

$$S_P = \mathcal{SN}(W_P \cdot \text{LN}(X_{patch})), \tag{23}$$

$$S_a = \mathcal{SN}(\text{Conv}(S_P)), \tag{24}$$

where L_P denotes the block length, Unfold represents the sliding window operation, and Conv employs group convolutions to enhance feature diversity.

Spiking text encoder processes short, medium, and long-term dependencies in text sequences through parallel processing by LIF neurons with varying time constants. For text embedding $X_t \in \mathbb{R}^{B\times L\times d}$, the encoding process is as follows:

$$S_f = \mathcal{SN}_{\tau=1.5}(W_f \cdot X_t), \tag{25}$$

$$S_m = \mathcal{SN}_{\tau=2.0}(W_m \cdot X_t), \tag{26}$$

$$S_s = \mathcal{SN}_{\tau=3.0}(W_s \cdot X_t), \tag{27}$$

where S_f, S_m, and S_s correspond to branches with different time constants. Feature fusion is achieved through concatenation and projection to obtain S_{fused}, and finally, feature information is integrated using residual connection:

$$S_{res} = \mathcal{SN}(W_{res} \cdot X_t), \tag{28}$$

$$S_t = \text{LN}(S_{fused} + S_{res}). \tag{29}$$

For SNNs methods, we uniformly employ the aforementioned encoding process to construct modality-specific spike encoders; for ANNs methods, we utilize pulse encoders with their architecturally equivalent counterparts within ANNs.

Multi-head Spiking Self Attention. We extend the Spike-Driven Self-Attention (SDSA) proposed in (Yao et al., 2023) to a multi-head architecture. Specifically, the floating-point matrices Q, K, and V are computed by three learnable linear matrices, followed by pulse neurons that transform Q, K, and V into spiking tensors Q_S, K_S, and V_S. SDSA is expressed as:

$$\text{SDSA}(Q, K, V) = \mathcal{SN}(\text{SUM}_c(Q_S \otimes K_S)) \otimes V_S, \tag{30}$$

where $\otimes$ denotes the Hadamard product, and SUM_c represents the sum of each column. The Hadamard product between spiking tensors corresponds to a mask operation (Yao et al., 2023).

In practice, we reshape Q, K, and V into multi-head formats. We then divide Q, K, and V into H segments and perform parallel H SDSA operations, referred to as H-head SDSA. The Multi-head Spiking Self-Attention (MSSA) in SpikeMoE is represented as follows:

$$Q = (q_1, \ldots, q_H), \tag{31}$$

$$K = (k_1, \ldots, k_H), \tag{32}$$

$$V = (v_1, \ldots, v_H), \tag{33}$$

$$\tilde{\text{M}} = \big[\text{SDSA}_1(q_1, k_1, v_1); \ldots; \text{SDSA}_1(q_1, k_1, v_1)\big], \tag{34}$$

$$\text{MSSA}(Q, K, V) = \text{Linear}\big(\tilde{\text{M}}\big). \tag{35}$$

Spiking Expert. Each of the K experts is a spike-based MLP. Given an input $X \in \mathbb{R}^{T\times B\times N\times C}$, the k-th expert computes:

$$\text{E} = \mathcal{SN}(\text{Linear}(\text{BN}(X))), \tag{36}$$

$$\text{Expert}_k(\text{E}) = \mathcal{SN}(\text{Linear}(\text{BN}(\text{E}))), \tag{37}$$

where BN is BatchNorm.

A.2.3 BASELINE METHODS

ANN Baselines. **ShaSpec** (Wang et al., 2023a) employs multiple loss functions to learn both modality-shared and modality-specific information, and utilizes the shared information to estimate missing modalities. **TF** (Zadeh et al., 2017) extracts modality interactions through multimodal sub-networks and a tensor fusion layer. **mmFormer** (Zhang et al., 2022), a typical multimodal Transformer, features multi-head self-attention mechanisms. **FuseMoE** (Han et al., 2024) respects missing modalities and jointly learns from both single and missing modality combinations of data inputs. In experiments, FuseMoE utilizes a softmax gate.

SNN Baselines. **Spikformer** (Zhou et al., 2022) pioneers the integration of Transformer architectures into SNNs by proposing Spiking Self-Attention. It significantly enhances the representation capacity of SNNs for visual feature extraction while maintaining biological interpretability. **Spike-driven Transformer** (Yao et al., 2023) introduces a purely event-driven attention paradigm. It effectively replaces all dense matrix multiplications in vanilla Transformers with sparse addition operations based on spike masks. **Spikingformer** (Zhou et al., 2023) further improves SNN-based Transformers by optimizing spike-driven residual learning and feature mapping mechanisms. **SpikeLM** (Xing et al., 2024) is an advanced spiking language model specifically designed for natural language processing (NLP) tasks. By adapting spike representations for text sequences, it demonstrates the effectiveness and scalability of SNNs on standard language understanding benchmarks such as GLUE. (Zhou et al., 2022) .**WeightAttention** (Liu et al., 2022) dynamically assigns weights to measure the importance of visual and auditory modalities, thereby effectively integrating information from these two modalities. **SCA** (Guo et al., 2023) integrates image and audio sub-modules and combines relative positional bias (RPB) to enhance the ability to capture information from different positions. **S-CMRL** (He et al., 2025) employs spatiotemporal spike attention to extract complementary features across modalities and combines cross-modal residual learning to enhance feature integration. Additionally, S-CMRL introduces a semantic alignment optimization mechanism to align cross-modal features within a shared semantic space, thereby improving their consistency and robustness.

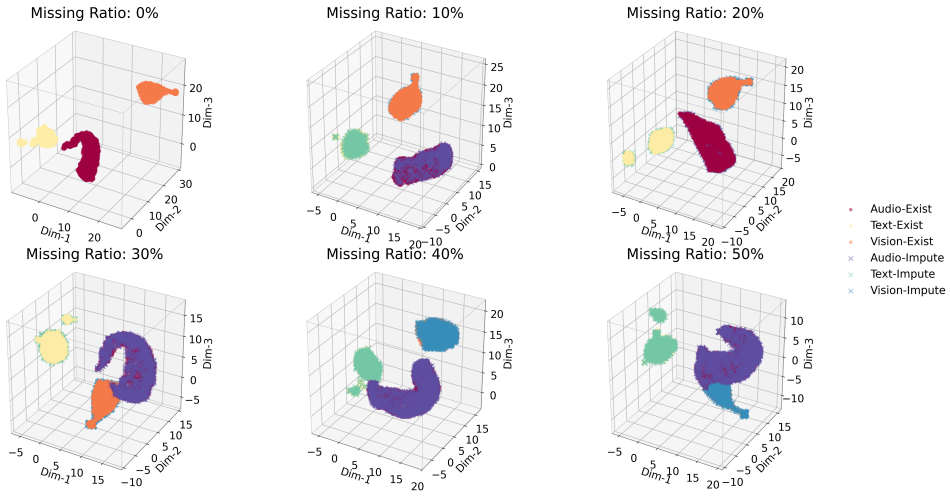


Figure 9: UMAP visualization of multi-modal embedding space: CMU-MOSEI.

A.2.4 COMPLEXITY AND ENERGY CONSUMPTION ANALYSIS

We compare the complexity of our proposed method and baseline models in terms of operations (OPs) and parameters, along with their AUC values on the CMU-MOSI dataset at a missing ratio of 50%. In SNNs, OPs refer to SOPs, while in ANNs they refer to FLOPs. SOPs are computed as follows:

$$\text{SOPs}(l) = fr \times T \times \text{FLOPs}(l), \tag{38}$$

where l denotes the block/layer within the Transformer architecture, fr represents the emission rate of the input pulse sequence for the block/layer, and T is the simulated time step for spike neuron. FLOPs(l) refers to the number of floating-point operations for l, specifically the count of

Table 6: Complexity and Energy Consumption analysis

	ANNs				SNNs			Ours
Metric	**ShaSpec**	**TF**	**mmFormer**	**FuseMoE**	**WeightAttention**	**SCA**	**S-CMRL**	**SpikeMoE**
Ops (M)	21.73	48.42	1577.63	199.8	47.07	32.57	289.60	**45.22**
Params	2,177,406	4,211,331	11,373,315	243,543,180	131,585	264,192	2,685,729	**2,472,370**
AUC	55.60±0.12	60.49±0.80	61.87±0.67	67.28±0.28	65.81±0.05	64.49±0.43	65.71±0.43	**67.30±0.71**
Energy (mJ)	20.13	25.19	40.39	11.71	10.38	16.95	14.67	**5.29**

multiply-accumulate (MAC) operations. SOPs denote the number of spike-based accumulate (AC) operations.

For all theoretical energy consumption calculations in this paper, we adopt the theoretical energy consumption estimation method proposed by (Yao et al., 2023). The results demonstrate that SpikeMoE achieves the optimal balance between complexity and performance. Detailed data for the complexity analysis are provided in Table 6. The results demonstrate that ShaSpec is the simplest model, yet its task performance is not comparable. In contrast, mmformer and S-CMRL are highly complex in terms of OPs and parameters, but their performance is not optimal. FuseMoE demonstrates performance comparable to our proposed method, yet it requires significantly more OPs and parameters. SpikeMoE is the optimal model that best balances complexity and performance.

A.2.5 Visualization of Experiment

In this section, we provide visual evidence to further support the effectiveness of our proposed missing modality modeling approach. We present the UMAP projection of multimodal embeddings on CMU-MOSEI in Figure 9. The visualization clearly demonstrates that the interpolated modalities form distributions consistent with their original counterparts. Combined with the results in Figure 5, this indicates that our proposed two-stage missing-modality modeling strategy preserves semantic alignment across modalities, proving its effectiveness in the missing-modality filling process.

A.2.6 Implementation Details and Hardware Setting

For the spiking neuron model, LIF neurons are configured with an initial membrane potential of 0 and a firing threshold of 0.5. To facilitate efficient backpropagation, a sigmoid function with parameter $\alpha = 4$ is employed as the surrogate gradient function, expressed as $Sigmoid(x) = 1/(1+\exp(-\alpha x))$. Key hyperparameters are listed in Table 7. SpikeMoE is optimized through the minimization of the expected risk based on the loss function $L_{balance}$. The load-balance loss $\mathcal{L}_{\text{balance}}$ encourages the equitable utilization of experts, preventing some from being overused while others are neglected. By integrating these losses, we achieve the overall objective:

$$\mathcal{L}_{\text{balance}} = \text{Var}(p) + \alpha(\log E - \text{Entropy}(\mathbf{p})), \tag{39}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{balance}}, \tag{40}$$

where λ is the coefficient of $\mathcal{L}_{\text{balance}}$. $\mathcal{L}_{\text{task}}$ is the cross-entropy loss.

A.2.7 Discussion of Sensitivity Analysis

To thoroughly evaluate the impact of hyperparameters on model performance, we conduct a sensitivity analysis on the CMU-MOSEI dataset (missing ratio=50%). We analyze the number of experts, number of expert layers, WTA-timesteps, and inhibition strength (Figure 8). The detailed discussion is as follows:

Model Capacity and Depth: As the number of experts increases from 4 to 8 and the expert layers from 2 to 5, both AUC and F1 metrics show a steady upward trend. This indicates that expanding the expert pool and deepening the MoE structure effectively enhance the network's ability to fit complex data distributions. Consequently, the model learns more fine-grained and hierarchically rich specialized features. This also demonstrates the promising scalability of SpikeMoE.

Temporal Dynamics: In the temporal dimension, the model achieves optimal performance when the WTA-timesteps are set to 14. If the time steps are too short, neurons cannot sufficiently accumulate membrane potential, depriving routing decisions of adequate temporal dynamics. Conversely, per-

Table 8: Experiment Setting I: Main results on CMU-MOSEI. Top-3 results are highlighted as first, second and third.

Missing Ratio	Metric	ANNs				SNNs			Ours
		ShaSpec (Wang et al., 2023a)	TF (Zadeh et al., 2017)	mmFormer (Zhang et al., 2022)	FuseMoE (Han et al., 2024)	Weight Attention (Liu et al., 2022)	SCA (Guo et al., 2023)	S-CMRL (He et al., 2025)	SpikeMoE
0%	F1	47.48±0.27	38.89±0.54	56.75±0.47	57.04±0.15	59.40±0.42	59.36±0.06	58.09±0.19	59.80±0.17
	AUC	67.93±0.56	64.26±0.61	76.42±0.24	78.54±0.42	73.59±0.08	74.41±0.13	72.63±0.21	78.63±0.53
10%	F1	47.41±0.18	34.94±1.10	55.37±0.47	55.90±0.55	54.97±0.35	56.81±0.35	56.23±0.32	56.83±0.23
	AUC	69.26±0.22	61.05±0.34	75.36±0.53	76.38±0.39	72.18±0.34	73.12±0.23	73.17±0.09	76.47±0.49
20%	F1	46.54±1.04	35.37±0.94	52.91±0.75	54.44±0.27	53.14±0.17	54.21±0.11	52.84±0.45	54.35±0.57
	AUC	68.05±0.56	59.54±0.22	73.02±0.65	74.32±0.16	69.92±0.12	70.32±0.54	69.95±0.02	74.92±0.86
30%	F1	45.05±0.90	35.36±0.84	49.71±0.30	52.27±0.35	51.60±0.55	51.78±0.31	50.51±0.03	54.22±0.49
	AUC	66.89±0.31	60.23±0.31	71.28±0.10	72.22±0.52	68.54±0.07	68.14±0.50	67.46±0.33	73.66±0.41
40%	F1	43.49±0.95	32.23±0.61	48.06±0.54	49.32±0.17	48.30±0.03	47.90±0.39	48.14±0.41	48.89±0.62
	AUC	65.32±0.59	60.09±0.35	69.31±0.10	69.85±0.11	67.40±0.47	63.98±0.12	66.75±0.36	69.92±0.68
50%	F1	42.75±1.22	33.01±0.51	47.39±0.16	43.08±0.47	46.50±0.54	46.49±0.26	47.44±0.24	49.82±0.74
	AUC	62.32±0.42	58.20±0.17	66.61±0.22	67.63±0.55	63.84±0.39	63.06±0.05	64.09±0.07	68.80±0.20

formance slightly declines when the time steps exceed 14. This is likely because excessively long integration introduces temporal noise without providing additional representation benefits.

Competition-Inhibition Balance: For the inhibition strength—a core parameter of the spiking gating network—the model peaks in performance at 0.3. This validates the importance of dynamically balanced selection. A moderate inhibition strength (0.3) successfully prevents expert collapse and ensures sparse activation. However, overly weak inhibition (< 0.3) reduces competition among experts, degrading routing sparsity. Conversely, excessively strong inhibition (> 0.3) causes over-suppression, making it difficult for certain expert neurons to activate and thereby limiting collaborative feature processing across multiple experts.

Table 7: Hyperparameter setting

Parameter	Value
Number of Expert	4 for ImageNet-1K and GLUE, 6 for CMU-MOSI and CMU-MOSEI, 4 for UrbanSound8k-AV
Top-K Expert	2
Timesteps T	4 for ImageNet-1K, GLUE and UrbanSound8K-AV; 6 for CMU-MOSI and CMU-MOSEI
Learning Rate	5e-5
Hidden Size	128
MoE layers	2
K for Sampling	10
wta-timesteps t	14
Refractory Period	2
Inhibition Strength	0.3
Training Epoch	300
Random Seed	42, 43, 44
Dropout	0.3

A.3 Results on CMU-MOSEI

We conducted experiments using experimental setup I and II on the CMU-MOSEI dataset, with results shown in Table 8 and 9.

Table 9: Experiment Setting II: Main results on CMU-MOSEI. Top-3 results are highlighted as first, second and third.

Testing Modality			Metric	ANNs				SNNs			Ours
Video	Text	Audio		ShaSpec (Wang et al., 2023a)	TF (Zadeh et al., 2017)	mmFormer (Zhang et al., 2022)	FuseMoE (Han et al., 2024)	Weight Attention (Liu et al., 2022)	SCA (Guo et al., 2023)	S-CMRL (He et al., 2025)	SpikeMoE
✓	✓		F1	40.75±0.14	31.38±0.61	56.02±0.56	55.85±0.69	33.34±0.28	52.80±0.76	55.62±0.34	56.25±0.32
			AUC	64.47±0.10	56.98±0.20	76.60±0.39	75.81±0.82	65.91±0.45	71.60±0.04	73.21±0.45	77.09±0.44
✓		✓	F1	38.15±0.92	31.73±0.16	37.12±0.66	49.30±0.67	22.00±0.49	29.01±0.04	22.10±0.20	49.50±0.36
			AUC	58.41±0.10	58.41±0.17	56.72±0.84	56.32±0.62	50.91±0.57	51.26±0.25	50.30±0.06	57.08±0.24
	✓	✓	F1	40.82±0.12	31.86±0.23	56.06±0.29	57.02±0.60	56.06±0.48	33.90±0.21	21.93±0.83	58.53±0.02
			AUC	64.46±0.14	58.64±0.13	73.91±0.04	73.42±0.60	73.31±0.10	56.54±0.17	50.29±0.32	75.47±0.19
✓			F1	40.37±0.46	31.59±0.23	55.07±0.35	54.02±2.88	32.65±0.44	50.80±0.61	53.41±0.29	55.09±0.42
			AUC	63.52±0.19	58.71±0.35	75.82±0.28	74.84±0.14	64.96±0.59	72.06±0.39	70.62±0.02	76.03±0.39
	✓		F1	42.16±0.22	30.84±0.26	55.89±0.37	55.87±0.68	33.20±0.68	51.04±0.03	52.32±1.25	55.62±0.25
			AUC	64.40±0.24	58.89±0.40	74.83±0.28	73.83±0.05	65.02±0.07	71.02±0.06	69.98±0.33	74.23±0.22
		✓	F1	42.45±0.85	30.71±0.71	55.06±0.29	55.64±1.65	49.06±0.65	52.01±0.51	22.93±0.35	55.93±0.47
			AUC	65.16±0.04	58.32±0.04	75.85±0.09	76.14±0.35	70.31±0.51	70.16±0.34	50.21±0.14	76.84±0.15

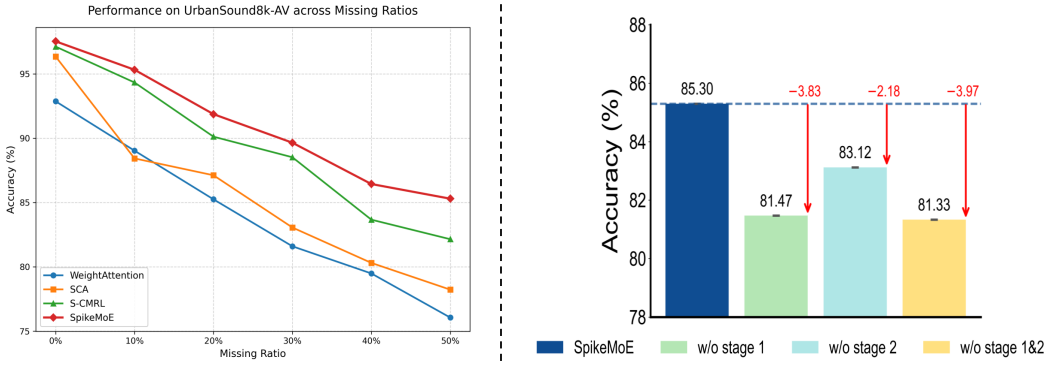
A.4 Results on UrbanSound8K-AV

A.4.1 Primary Results

We conducted experiments using experimental setup I on the UrbanSound8k-AV dataset, with results shown in the Figure 10. The experimental results demonstrate that SpikeMoE achieves state-of-the-art performance when compared against all SNN baseline methods.

A.4.2 Ablation Study

We conducted ablation studies on UrbanSound8k-AV as outlined in Fig. 10. We sequentially discarded the two stages of the missing modality modeling strategy and replaced missing modalities with zeros. When both stages were removed, Acc decreased by 3.97% on UrbanSound8k-AV. Additionally, we observed the same conclusion on CMU-MOSI and CMU-MOSEI: removing stage 1 resulted in a greater decline than removing stage 2. This suggests that in missing modal modeling, the modal interactions captured by the modal pool may play a more critical role.



(a) (b)

Figure 10: (a) Performance on UrbanSound8k-AV across Missing Ratios; (b) Ablation study on UrbanSound8k-AV (Experiment Setting I, Missing Ratio 50%).

A.5 Inference Latency and Throughput Analysis

We evaluated the inference latency of SpikeMoE on a single NVIDIA RTX 4090 GPU (Batch Size = 64). As shown in Fig.11, the left panel demonstrates that as the number of experts scales from 4 to 8, the computational overhead of the k-WTA router remains exceptionally low (accounting for only 3.7% to 4.4% of the total latency). Meanwhile, the backbone and expert layer latencies remain strictly constant due to the sparse Top-K activation. The right panel indicates that increasing the wta-timesteps (t) from 8 to 20 yields only a marginal increase in router execution time ($\sim$1 ms). Consequently, the total inference latency remains stable, confirming that our routing mechanism scales efficiently without introducing computational bottlenecks.

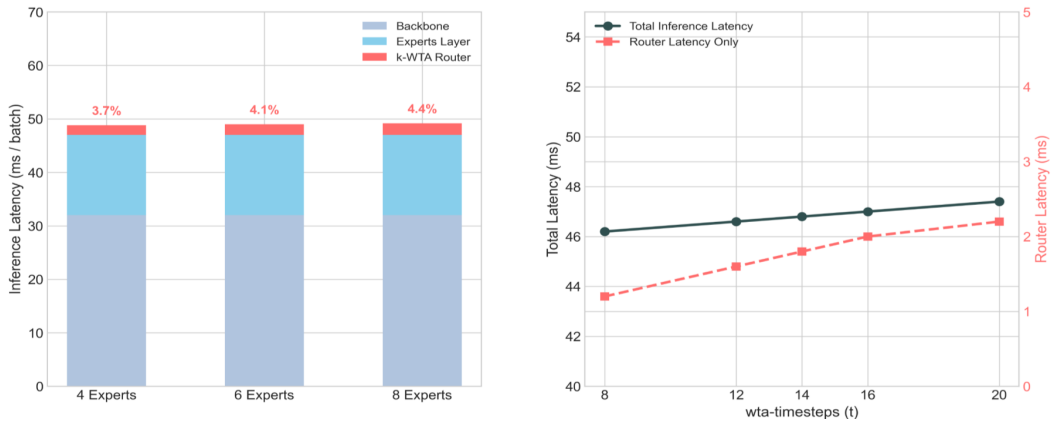


Figure 11: (a) Latency Breakdown across Expert Counts; (b) Impact of wta-timesteps on Latency.

A.6 Algorithm Workflow

In this section, we present the algorithmic flow and pseudocode for the k-WTA Router (Algorithm 1).

Algorithm 1 k-WTA Router Algorithm

Require: Input $X \in \mathbb{R}^{T\times B\times N\times C}$, experts E, wta-timesteps t_{max}, time const τ, inhibit γ, eps ε
Ensure: Routing weights $w_{\text{expert}} \in \mathbb{R}^{B\times E}$, Selected experts $\mathcal{E}_{\text{selected}}$

Initialize: $V \leftarrow 0$, $R \leftarrow 0$, $c \leftarrow 0$, Weights L_{inhibit}, Mask M_{inhibit}
for $t = 1$ **to** t_{max} **do** ▷ **Phase 1: Spiking & Lateral Inhibition**
  $V[t] \leftarrow V[t-1] \cdot (1 - \frac{1}{\tau}) + I[t]$
  $V_{\text{inhibit}}[t] \leftarrow V[t] - S[t-1] \cdot (L_{\text{inhibit}} \odot M_{\text{inhibit}}) \cdot \gamma$
  for $b \in [1, B], e \in [1, E]$ **do**
    if $V_{\text{inhibit}}[b, e, t] \geq V_{th}[t]$ **and** $R[b, e, t] \leq 0$ **then**
      $S[b, e, t] \leftarrow 1$
      $V[b, e, t] \leftarrow 0$; $R[b, e, t] \leftarrow t_{\text{refra}}$ ▷ Membrane reset & refractory
      $c[b, e] \leftarrow c[b, e] + 1$ ▷ Accumulate spikes
    else
      $S[b, e, t] \leftarrow 0$
    end if
  end for
  $R[t+1] \leftarrow \max(0, R[t] - 1)$ ▷ Update refractory counter
end for
Expert Selection: ▷ **Phase 2: Routing Assignment**
$\mathcal{E}_{\text{selected}} \leftarrow \text{TopK}_{i\in[E]}(c_i)$ ▷ Select top-K experts based on spike count
Compute Weights: $\forall b \in [1, B], \forall e \in [1, E]$

$$w[b, e] \leftarrow \begin{cases} \frac{c[b,e]+\varepsilon}{\sum_{j\in\mathcal{E}_{\text{selected}}}(c[b,j]+\varepsilon)} & \text{if } e \in \mathcal{E}_{\text{selected}} \\ 0 & \text{otherwise} \end{cases}$$

return w_{expert}, $\mathcal{E}_{\text{selected}}$

A.7 Limitations

Limitations of this study include the need to test the performance of SpikeMoE when scaled up to larger models and more modalities, as well as limitations related to hardware deployment. These issues will be addressed in future research. In the future, we will evaluate SpikeMoE's performance at larger scales and explore deploying SpikeMoE on hardware platforms such as Field-Programmable Gate Arrays (FPGAs) to assess its practical performance. In this process, we will optimize appropriate read-write data flows and memory access schemes to enhance the model's inference speed.