WHICH INTERNAL NEURAL REPRESENTATIONS SHOULD BECOME SPIKES? TASK-RELEVANT SPATIOTEMPORAL THRESHOLD ADAPTATION FOR AUDIO-VISUAL SPIKING NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spiking neural networks (SNNs), regarded as the third generation of neural networks, exhibit rich spatiotemporal dynamics. Audio-visual SNNs (AV-SNNs) can exploit the temporal nature of spike-based computation and event-driven processing for efficient multimodal learning. However, existing AV-SNN studies often leave a more fundamental question implicit: which internal neural representations should become spikes? Different modalities exhibit different task-relevant spatiotemporal structures, making it necessary to preserve highly task-relevant representations while distinguishing redundant ones. We propose **Spatio-Temporal Modality-Adaptive Thresholding (ST-MAT)**, consisting of three components: **Task-Relevant Spatiotemporal Support Estimation (TRSSE)** estimates task relevance from continuous pre-spike representations and identifies highly task-relevant locations in the spatiotemporal domain. **Semantic–Redundancy Balancing (SRB)** measures how task-relevant information is distributed over non-redundant spatiotemporal locations and balances semantic retention against redundancy rejection. **Slow Bounded Threshold Adaptation (SBTA)** learns neuronal thresholds through slow bounded updates to determine which continuous responses should be converted into spikes. Rather than prescribing how many spikes a modality should emit, ST-MAT changes which internal neuronal representations are converted into spikes. ST-MAT achieves 78.47% on AVE, 79.44% on CREMA-D, and 99.54% on UrbanSound8K-AV, improving ST-MAT with fixed threshold by 5.2%, 3.23%, and 1.14%, respectively. It also surpasses the same-architecture ANN on AVE and CREMA-D. Visualizations and ablation experiments show that ST-MAT converts high-task-value internal neuronal representations into spikes while reducing redundancy, thereby substantially improving AV-SNN performance. ST-MAT introduces no additional inference overhead, making it well suited to brain-inspired edge multimodal fusion.

1 INTRODUCTION

Audio-visual perception requires learning from sensory streams with different temporal and spatial organizations. Spiking neural networks (SNNs) provide rich spatiotemporal dynamics by converting membrane-potential information into discrete spikes over successive timesteps, enabling efficient processing of spatiotemporal information with substantial energy-saving potential. This dynamic, event-driven representation is particularly attractive for multisensory integration and makes SNNs well suited to audio-visual fusion. Existing audio-visual SNNs (AV-SNNs) mainly improve fusion by redesigning how already generated audio and visual spikes interact, including cross-modal attention, current-level integration, semantic alignment, temporal attention, and hierarchical binary attention (Liu et al., 2022; Guo et al., 2024; Jiang et al., 2023; He et al., 2025; Shen et al., 2025; Saleh and Ahmadi, 2026). These studies generally assume that a more fundamental process has already been handled correctly, namely that continuous internal neuronal representations have been appropriately discretized into spike representations. A spike, however, is not the original internal information; it is the discrete outcome of thresholding the membrane potential. Before asking how audio and

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.

visual spikes should be fused, we therefore ask a more fundamental question: which internal neuronal representations should become spikes?

Our experiments reveal that spike firing density is not tightly coupled with recognition quality across different modalities and tasks, as shown in Fig. 1. Neither increasing nor decreasing the number of emitted spikes provides a reliable objective for improving SNN representations. Insufficient firing may fail to transmit enough task-relevant information, whereas excessive firing can introduce redundant representations that interfere with effective audio–visual fusion. The central question is therefore not how many spikes a modality should generate, but which components of the current internal neuronal representation carry task-discriminative information and which mainly repeat already available spatiotemporal information. Selecting modality-relevant internal representations while reducing redundant spike activations is thus critical to improving audio–visual fusion in SNNs.

When external information reaches a spiking neuron, it drives the membrane potential to evolve over time, and a spike is emitted once the membrane potential crosses the firing threshold. The threshold therefore determines which internal neuronal representations are exposed to subsequent computation as discrete spikes. A fixed default threshold implicitly assumes that heterogeneous modalities can share the same internal-representation selection boundary, even though audio and visual streams have different membrane-potential distributions and different temporal and spatial structures. An effective threshold should satisfy two objectives simultaneously: preserve representations that are relevant to the task and suppress representations that are largely redundant. This raises two questions that are central to the continuous-to-spike conversion process: how should the task value of an internal neuronal representation be defined, and how can redundant representations be identified in the joint spatiotemporal representation space of an SNN?

To address this problem, we propose **Spatio-Temporal Modality-Adaptive Thresholding (ST-MAT)**, whose core selection process uses two coupled mechanisms to preserve task-relevant internal responses while suppressing redundant ones. Neural responses in multisensory systems depend on when and where signals occur and on their behavioral relevance (Stein and Stanford, 2008; Rohe and Noppeney, 2016; Senkowski and Engel, 2024), which motivates **Task-Relevant Spatiotemporal Support Estimation (TRSSE)**. TRSSE first assigns each pre-spike response a task-relevance score from the gradient-weighted classification margin, then measures how broadly task-relevant activity is supported across distinct spatiotemporal locations after correlated repetitions are discounted; the resulting bounded support score separates useful distributed structure from firing density alone. Inspired by selective redundancy control in multisensory processing, we further introduce **Semantic–Redundancy Balancing (SRB)**. SRB combines local task relevance with correlation-based redundancy exposure to construct semantic-retention and redundancy-rejection distributions, and selects a threshold that balances the two; this avoids both under-activation that discards useful responses and over-activation that repeatedly transmits the same information. Finally, we employ **Slow Bounded Threshold Adaptation (SBTA)** to update the threshold slowly, preventing overly rapid changes in the spike distribution from destabilizing the data used to select the next threshold. Together, ST-MAT establishes a new paradigm for efficient multimodal integration in SNNs: internal neural representations are selectively converted into spikes according to their task relevance, preserving high-value representations while suppressing redundant activation.

The contributions are summarized as follows.

(1) We find that the number of emitted spikes cannot directly improve AV-SNN performance. We argue that the key is to convert internal neuronal representations with high task value and low redundancy into spikes.

(2) We propose ST-MAT, which consists of three components: TRSSE, SRB, and SBTA. TRSSE estimates the task relevance of continuous internal neuronal representations from modality-specific spatiotemporal contributions to classification and measures their effective support over distinct spatiotemporal locations, helping preserve high-value representations. SRB combines local redundancy with task contribution to balance semantic retention and redundancy rejection, thereby suppressing repeated low-value responses while retaining useful information. SBTA derives the target threshold from task value and redundancy and updates it slowly with bounded steps, enabling stable conversion of high-task-value, low-redundancy internal neuronal representations into spikes.

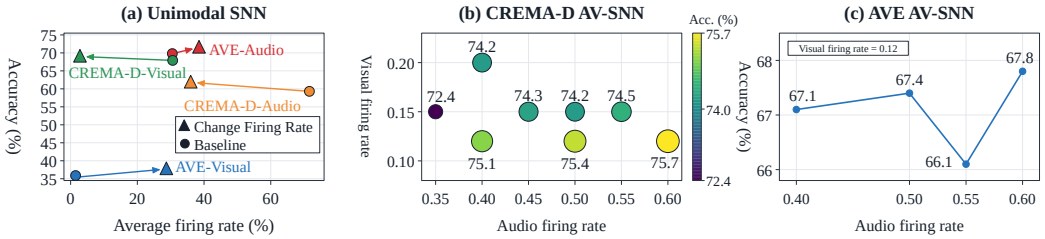

Figure 1: Spike firing rate cannot reflect AV-SNN accuracy.

(3) We conduct comparative experiments and ablation studies on AVE, CREMA-D, and UrbanSound8K-AV. ST-MAT achieves accuracies of 78.47%, 79.44%, and 99.54%, respectively, demonstrating its effectiveness and generalization across audio-visual tasks. The ablation studies and visualization results further show that task-relevance modeling and semantic–redundancy balancing enable high-value internal neuronal representations to be converted into external spike activations while reducing redundant information.

2 Related Work

2.1 Audio-Visual Spiking Neural Networks

Early multimodal SNNs combine modality-specific spikes through attention or current-level interaction. WeightAttention introduces attention-guided event-based multimodal learning (Liu et al., 2022), SMMT/SCA uses transformer-style spiking cross-attention (Guo et al., 2024), and CMCI integrates complementary current information across modalities (Jiang et al., 2023). More recent systems enrich the fusion operator with semantic alignment and residual exchange (He et al., 2025), multimodal interactive spiking units (Liu et al., 2025), temporal attention and adaptive fusion (Shen et al., 2025), hierarchical bidirectional binary QK interaction (Saleh and Ahmadi, 2026), and compact or brain-inspired fusion mechanisms (He et al., 2026; Oikonomou et al., 2026; Zhong et al., 2026). These studies mainly optimize how already-formed spike representations interact.

2.2 Adaptive Neuronal Dynamics and Intrinsic Excitability

Adaptive neuronal parameters can improve temporal expressivity in SNNs. Spike-frequency adaptation provides long-timescale memory (Bellec et al., 2018), learnable membrane constants show that neuronal dynamics need not remain globally fixed (Fang et al., 2021), and surrogate-gradient learning enables end-to-end optimization (Neftci et al., 2019). Most trainable-neuron approaches optimize membrane parameters directly with task gradients or local firing history. ST-MAT differs by using a task-structured target for the firing threshold and by updating that threshold on a slower bounded timescale.

2.3 Multisensory Spatiotemporal Integration

Multisensory responses depend on temporal and spatial relationships across sensory inputs and on the reliability and behavioral relevance of each modality (Stein and Stanford, 2008; Rohe and Noppeney, 2016). Multi-timescale dynamics also contribute to sensory integration and segregation (Senkowski and Engel, 2024). ST-MAT does not attempt to model these biological circuits directly. It adopts the more limited computational principle that the continuous-to-spike conversion boundary should depend on the structure and task relevance of ongoing multimodal activity.

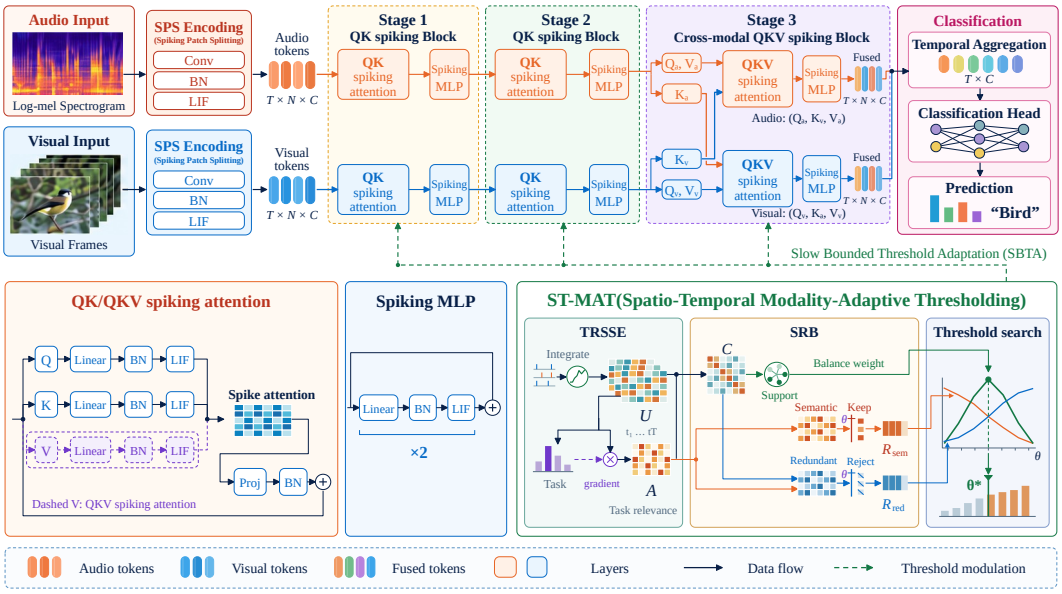

Figure 2: ST-MAT architecture.

3 Preliminaries and Problem Formulation

3.1 Problem Formulation

For modality $m \in \{A, V\}$ and stage l, let $U_{m,l} \in \mathbb{R}^{T \times N_l \times D_l}$ denote the continuous pre-spike representation produced by the neuron dynamics. The corresponding external spike representation is

$$S_{m,l} = H(U_{m,l} - \theta_{m,l}). \tag{1}$$

The threshold $\theta_{m,l}$ therefore controls which values in the continuous internal representation are exposed to the next computation as discrete spikes. A global firing rate describes only the number of emitted events. It does not tell us whether the transmitted responses are useful for the task or largely repeated. We seek a threshold that uses the task-related spatiotemporal organization of $U_{m,l}$ to preserve informative responses and suppress redundant ones without enforcing a target firing rate.

3.2 Audio-Visual SNN Backbone

Figure 2 shows the network used throughout the paper together with ST-MAT. We use LIF (Leaky Integrate-and-Fire) neurons, and audio log-mel features and sampled video frames are encoded by SPS (Spiking Patch Splitting; `Conv+BN+LIF`) into T-step spike-token sequences. Stages 1 and 2 process the two modalities independently with QK spiking-attention blocks and spiking MLPs. Stage 3 introduces cross-modal QKV interaction, after which the fused spike tokens are temporally aggregated for classification. The lower part of Fig. 2 shows the QK/QKV spiking-attention block, the spiking MLP, and the ST-MAT pathway. ST-MAT reads the continuous pre-spike representation, estimates task relevance and non-redundant support, and adjusts the threshold before the next spike representation is formed.

4 Spatio-Temporal Modality-Adaptive Thresholding

The experiments in Fig. 1 expose three difficulties. Firing rate does not indicate which responses are useful, high task relevance can still be duplicated across correlated locations, and an aggressively updated threshold can destabilize the representation that produced it. ST-MAT addresses these difficulties in the same order. ST-MAT comprises Task-Relevant Spatiotemporal Support Estimation (TRSSE), Semantic–Redundancy Balancing (SRB), and Slow Bounded Threshold Adaptation (SBTA).

TRSSE estimates the task relevance of internal neuronal responses and measures how broadly useful information is distributed over distinct spatiotemporal locations. SRB combines task relevance with local redundancy to derive a threshold objective that preserves semantic information while suppressing repeated low-value responses. SBTA then moves the deployed threshold toward the preferred value on a stable timescale, allowing the conversion boundary to adapt without abruptly changing the spike distribution.

4.1 Task-Relevant Spatiotemporal Support Estimation (TRSSE)

A firing threshold is useful only if the method can distinguish responses that matter to the current decision from responses that merely increase activity. Therefore, we compute task relevance on the continuous pre-spike representation before the current neuron emits a spike. Let b index a sample in the calibration minibatch, t the SNN timestep, n the token or spatial position, and d the feature/channel dimension. For a fixed modality m and stage l, a lightweight modality-specific branch readout produces class logits $\mathbf{f}_{m,b} = [f_{m,b,1}, \ldots, f_{m,b,K}] \in \mathbb{R}^K$. Here $f_{m,b,c}$ is the logit of class c, y_b is the ground-truth class of sample b, and K is the number of classes. We define the classification margin as $M_{m,b} = f_{m,b,y_b} - \max_{c \neq y_b} f_{m,b,c}$ and denote $A_{m,l,b,t,n,d}$ as the first-order task relevance of the continuous pre-spike representation $U_{m,l,b,t,n,d}$ to the classification margin, computed as:

$$A_{m,l,b,t,n,d} = \left| U_{m,l,b,t,n,d} \frac{\partial M_{m,b}}{\partial U_{m,l,b,t,n,d}} \right|. \tag{2}$$

Task relevance alone does not resolve repeated information. Two locations can both be highly relevant while carrying nearly the same response pattern. We use site $i = (t, n)$, aggregate channel relevance as $a_i = \sum_d A_{i,d}$, and normalize it with $w_i = a_i / (\sum_j a_j + \varepsilon)$. Channel-specific relevance is $\pi_{i,d} = A_{i,d} / (\sum_c A_{i,c} + \varepsilon)$. The task-aware descriptor $Z_{i,d} = \sqrt{\pi_{i,d}}\, U_{i,d} / (\|U_i\|_2 + \varepsilon)$ retains the continuous response while emphasizing channels that matter to the decision. After flattening time and position, similarity between two sites is $C_{ij} = |\cos(Z_i, Z_j)|$. We then define:

$$\mathcal{R}_{\mathrm{ST}} = \sum_{i,j} w_i w_j C_{ij}^2, \qquad d_{\mathrm{ST}} = \frac{1}{\mathcal{R}_{\mathrm{ST}} + \varepsilon}, \qquad s_{\mathrm{ST}} = \frac{d_{\mathrm{ST}} - 1}{T \cdot N_l - 1}. \tag{3}$$

Here, $\mathcal{R}_{\mathrm{ST}}$ is the task-weighted global redundancy, measuring the pairwise similarity among task-relevant spatiotemporal sites; d_{ST} is the corresponding effective number of distinct task-relevant sites; and s_{ST} is the normalized task-relevant spatiotemporal support. We define s_{ST} to describe how broadly task-relevant information is distributed across distinct spatiotemporal locations. We prove the boundedness of s_{ST} and its non-equivalence to firing rate in Appendix A.

4.2 Semantic–Redundancy Balancing (SRB)

TRSSE tells us how broadly useful information is distributed, but the threshold still needs a concrete keep-or-suppress criterion. We define the local redundancy exposure of site i as $\xi_i = \sum_j w_j C_{ij}^2$, which satisfies $\mathcal{R}_{\mathrm{ST}} = \sum_i w_i \xi_i$. The normalized semantic mass is $p_{i,d}^{+} = A_{i,d} / (\sum_{j,c} A_{j,c} + \varepsilon)$. A response should not be suppressed simply because it is correlated with another response, since correlated locations can still be important. We therefore construct suppressible redundancy as $\tilde{p}_{i,d}^{-} = \xi_i (1 - \pi_{i,d})$ and normalize it to $p_{i,d}^{-} = \tilde{p}_{i,d}^{-} / (\sum_{j,c} \tilde{p}_{j,c}^{-} + \varepsilon)$. Redundancy receives large weight only where repetition is high and local task relevance is low.

Global support controls the preservation–suppression balance through $\alpha(s_{\mathrm{ST}}) = (1 + s_{\mathrm{ST}})/2$. For a candidate threshold θ, let $g_{i,d}(\theta) = \mathbb{I}[U_{i,d} \geq \theta]$. Semantic retention R_{sem} and redundancy rejection R_{red} are:

$$R_{\mathrm{sem}}(\theta) = \sum_{i,d} p_{i,d}^{+} g_{i,d}(\theta), \qquad R_{\mathrm{red}}(\theta) = \sum_{i,d} p_{i,d}^{-} [1 - g_{i,d}(\theta)]. \tag{4}$$

We define $J(\theta)$ as the semantic–redundancy balancing utility of candidate threshold θ:

$$J(\theta) = R_{\mathrm{sem}}(\theta)^{\alpha} R_{\mathrm{red}}(\theta)^{1-\alpha}. \tag{5}$$

We use a multiplicative form to jointly account for semantic retention R_{sem} and redundancy rejection R_{red}. If either objective becomes too extreme, the other factor approaches zero because retaining all

semantic information also preserves redundant information, whereas optimizing only for redundancy rejection can discard task-relevant semantics. This prevents one objective from freely compensating for failure of the other and converts the support structure from TRSSE into a threshold that selectively transforms useful continuous responses into spikes. When a non-trivial interior threshold exists, the all-pass and all-reject solutions are suboptimal. We prove this theoretically in Appendix A and further show how s_{ST} changes the preference among candidate thresholds.

4.3 Slow Bounded Threshold Adaptation (SBTA)

The threshold preferred by SRB is estimated from the current network representation. Replacing the deployed threshold with that value in one step would immediately redistribute spikes and invalidate part of the statistics used to select it. We therefore search only over observed continuous response values, with $\theta^*_{m,l} = \arg\max_{\theta \in \mathrm{sort}\{U_{i,d}\}} J_{m,l}(\theta)$, and treat $\theta^*_{m,l}$ as a target rather than an instantaneous replacement. Here, q denotes the index of the periodic threshold update, and the resulting threshold after this bounded update is denoted by θ^{q+1}. The deployed threshold follows:

$$\theta^{q+1} = \Pi_{[\theta_{\min},\theta_{\max}]} \left\{ \theta^q + \mathrm{clip}[\eta_\theta(\theta^* - \theta^q), -\Delta_\theta, +\Delta_\theta] \right\}. \tag{6}$$

In Eq. (6), θ^q is the current threshold at update index q, θ^* is the target threshold preferred by SRB, η_θ controls the update rate, Δ_θ is the maximum allowed change in one update, $\mathrm{clip}(\cdot)$ bounds the update step, and $\Pi_{[\theta_{\min},\theta_{\max}]}$ projects the threshold into the prescribed safe interval. Ordinary network weights are updated every minibatch, whereas our threshold update is periodic. When clipping and projection are inactive, the update contracts toward a fixed target, as proved in detail in Appendix A. At inference, only the learned thresholds are retained. Therefore, the deployed forward graph is identical to ST-MAT with fixed threshold. Algorithm 1 summarizes the complete workflow.

Algorithm 1 ST-MAT threshold update strategy

Require: Training mini-batch $\{(X_A, X_V, y)\}$, AV-SNN backbone, current thresholds $\{\theta_{m,l}\}$
for each training iteration **do**
  Encode audio and visual inputs with SPS and direct temporal replication
  Forward the AV-SNN using current thresholds to obtain membrane tensors $U_{m,l}$ and logits
  Update ordinary network weights with the task loss by surrogate-gradient backpropagation
  if iteration is a threshold-update step **then**
    for each modality $m \in \{A, V\}$ and selected stage l **do**
      Compute branch margin M_m and relevance map $A_{m,l}$ via Eq. (2)
      Run TRSSE to form w_i, Z_i, C and estimate s_{ST} with Eq. (3)
      Run SRB to derive ξ_i, p^+, p^- and maximize Eq. (5)
      Search observed continuous response values to obtain the preferred threshold $\theta^*_{m,l}$
      Update the deployed threshold with the slow bounded rule in Eq. (6)
    end for
  end if
end for
Inference fixes the learned thresholds

5 Experiments

5.1 Experimental Setup

We evaluate our method and ablation experiments on AVE (Tian et al., 2018), CREMA-D (Cao et al., 2014), and UrbanSound8K-AV using the same AV-SNN pipeline. AVE contains 4,143 ten-second videos from 28 audiovisual event categories, and CREMA-D contains 7,442 clips from 91 actors and six emotion categories. UrbanSound8K-AV contains 8,732 audio clips paired with corresponding visual samples across 10 urban-sound categories. Our network is the QK/QK/QKV AV-SNN shown in Fig. 2 with $T = 6$ simulation steps. SPS encoding produces audio and visual spike-token sequences, Stages 1–2 remain modality specific, and Stage 3 performs cross-modal QKV interaction. ST-MAT updates thresholds during training and fixes them at inference. At deployment, ST-MAT with fixed threshold and ST-MAT share the same forward graph. Appendix C provides further training details, including preprocessing and optimization.

5.2 Main Results

Table 1 presents our experimental results on the three datasets. On AVE, ST-MAT achieves 78.47% accuracy, substantially improving upon the 73.27% obtained by ST-MAT with fixed threshold. This indicates that our method can convert more high-value internal neuronal representations into external spike representations, thereby improving audio–visual fusion. ST-MAT also outperforms ST-MAT with learnable threshold, which achieves 77.48%, suggesting that finer-grained task-relevance modeling and redundancy discrimination help the model learn a more appropriate threshold. The results on CREMA-D and UrbanSound8K-AV support the same conclusion. In particular, on the more challenging CREMA-D dataset, ST-MAT reaches 79.44%, improving the fixed-threshold result of 76.21% and also exceeding the learnable-threshold result of 72.98%. On UrbanSound8K-AV, ST-MAT improves the fixed-threshold accuracy from 98.40% to 99.54%, approaching 100%, and outperforms the 99.20% achieved by the learnable-threshold variant. Directly making neuronal thresholds learnable can improve fusion performance to some extent, whereas ST-MAT uses finer-grained task-relevance estimation and redundancy discrimination to guide the model toward more suitable thresholds.

We compare ST-MAT with other publicly reported SNN methods on all three datasets. On AVE, ST-MAT achieves 78.47%, attaining the best accuracy among the SNN methods reported in Table 1 and substantially outperforming the other models. On CREMA-D, ST-MAT also shows strong performance with 79.44% accuracy, second only to EAVS-TF, which uses pretrained initialization. ST-MAT remains highly competitive on UrbanSound8K-AV, reaching 99.54%. Because performance on UrbanSound8K-AV is already close to 100%, the remaining room for improvement is very limited, making small absolute differences less informative for reflecting the effectiveness of ST-MAT.

We further evaluate an ANN model with the same network architecture as ST-MAT on all three datasets. It achieves 76.88% on CREMA-D, 77.11% on AVE, and 99.88% on UrbanSound8K-AV. The same-architecture ANN outperforms the listed ANN baselines on AVE and UrbanSound8K-AV and remains competitive on CREMA-D, indicating that the backbone itself is strong. Because SNNs communicate through sparse binary 0/1 spikes, this discretization can ordinarily constrain representational capacity relative to a continuous-valued ANN with the same architecture. Nevertheless, ST-MAT clearly surpasses the same-architecture ANN on CREMA-D and AVE. We attribute this to the task-relevance estimation and redundancy discrimination in ST-MAT explicitly exploiting the temporal dimension of SNN states, which provides additional spatiotemporal information that is unavailable to the ANN counterpart.

5.3 Valuable Neural Representation Selection

We visualize the distributions of several variables in the audio and visual modalities, as shown in Fig. 3, where the top and bottom rows correspond to the audio and visual branches, respectively. Figure 3(a) shows the membrane magnitude U in Eq. (1), which represents the magnitude distribution of the continuous pre-spike neuronal responses before thresholding. Both modalities mainly emphasize the central region, although their strongest responses occur at slightly different locations. Figure 3(b) shows the task-relevance variable A in Eq. (2), which measures the contribution of each continuous pre-spike response to classification. Compared with audio, the visual branch contains little valid image content near the top and bottom boundaries, so A is more concentrated in the central rectangular region. Figure 3(c) shows the local redundancy exposure ξ defined in Sec. 4.2, where a larger value indicates that the response at a spatiotemporal position is more strongly repeated by other task-relevant positions. Figure 3(d) shows the firing-improvement variable Δp. For the same continuous pre-spike response U, we compute the soft activation propensity under a threshold θ as $p_\theta(U) = \sigma((U-\theta)/\beta)$, where $\sigma(\cdot)$ is the sigmoid function and β is the temperature. We then calculate $\Delta p = p_{\theta^*}(U) - p_{\theta_0}(U)$ using the learned threshold θ^* and the original fixed threshold θ_0. Thus, Δp is not a difference in binary firing rate; it measures how threshold learning changes the tendency of the same internal response to become a spike, with larger positive values indicating a stronger increase in activation propensity. Comparing the spatial trends of (a)–(d) across the audio and visual modalities provides an intuitive visualization that ST-MAT preferentially converts internal neuronal representations with high task relevance and low redundancy into spikes.

Table 1: Accuracy comparison on CREMA-D, AVE, and UrbanSound8K-AV.

Model	Method	Pretrained	CREMA-D	AVE	Us8k-AV
ANN	OGM-GE (Concat.) (Peng et al., 2022)	✗	61.90	65.42	98.97
	OGM-GE (Sum.) (Peng et al., 2022)	✗	62.20	66.42	98.97
	OGM-GE (FiLM) (Peng et al., 2022)	✗	55.60	64.43	98.74
	OGM-GE (Gated) (Peng et al., 2022)	✗	60.62	62.70	99.08
	PMR (Concat.) (Fan et al., 2023)	✗	61.10	67.10	99.66
	PMR (Sum.) (Fan et al., 2023)	✗	59.40	68.10	99.77
	PMR (FiLM) (Fan et al., 2023)	✗	61.80	66.40	99.66
	PMR (Gated) (Fan et al., 2023)	✗	59.90	62.70	99.77
	AGM (Li et al., 2023)	✗	70.16	-	-
	QMF (Zhang et al., 2023)	✗	63.71	-	-
	SCAF-Net (ANN) (Ma et al., 2026a)	✗	80.83	-	-
	Same-architecture ANN (ours)	✗	**76.88**	**77.11**	**99.88**
SNN	SMMT (Guo et al., 2024)	✗	62.50	51.78	96.85
	WeightAttention (Liu et al., 2022)	✗	64.78	65.92	98.17
	SCA (Guo et al., 2024)	✗	66.53	60.20	97.83
	CMCI (Jiang et al., 2023)	✗	70.02	67.91	97.90
	S-CMRL (He et al., 2025)	✗	73.25	68.15	98.13
	MISNet-XS (Liu et al., 2025)	✗	66.45	58.44	97.71
	MISNet-S (Liu et al., 2025)	✗	68.72	65.31	98.09
	MISNet-L (Liu et al., 2025)	✗	75.22	67.24	98.52
	MISNet-XL (Liu et al., 2025)	✗	77.14	68.04	98.96
	TAAF-SNNs (Concat.) (Shen et al., 2025)	✗	77.55	70.65	-
	TAAF-SNNs (Sum.) (Shen et al., 2025)	✗	76.75	68.91	-
	IEMF + Spatial CA (He et al., 2026)	✗	73.11	-	-
	IEMF + ST CA (He et al., 2026)	✗	72.98	-	-
	SMLP (Oikonomou et al., 2026)	✗	-	-	97.20
	Spikceiver (Zhong et al., 2026)	✗	-	-	95.05
	SNNergy (Saleh and Ahmadi, 2026)	✗	78.38	72.14	<u>99.66</u>
	SCAF-Net (Ma et al., 2026a)	✗	79.03	-	-
	EAVS-TF (Ma et al., 2026b)	✓	<u>81.48</u>	-	-
	ST-MAT with fixed threshold	✗	76.21	73.27	98.40
	ST-MAT with learnable threshold	✗	72.98	77.48	99.20
	ST-MAT (Ours)	✗	**79.44**	**78.47**	**99.54**

5.4 Ablation Study

We conduct ablation experiments on AVE, with the results summarized in Table 2. Removing TRSSE, shown in the second row, reduces accuracy to 75.25%, which is 3.22% below ST-MAT. This indicates that threshold selection benefits from characterizing where task-relevant information is distributed. TRSSE helps convert higher-value internal neuronal representations into external spikes, thereby improving audiovisual fusion. Removing SRB, shown in the third row, gives 75.50%, a 2.97% decrease from ST-MAT. This shows that task relevance alone is insufficient when correlated responses repeatedly encode the same content. SRB helps the model distinguish redundant information and thereby improves fusion. Removing both TRSSE and SRB, shown in the fourth row, lowers accuracy to 74.01%, a 4.46% decrease, and causes a larger degradation than removing either TRSSE or SRB alone. This confirms that TRSSE and SRB are complementary. Task-relevance estimation by TRSSE can still retain some redundant responses, while SRB explicitly suppresses this redundancy, so their combination further improves fusion. Replacing the multiplicative utility in SBTA with a linear additive form, shown in the fifth row, yields 73.27%, which is 5.20% below ST-MAT. The additive

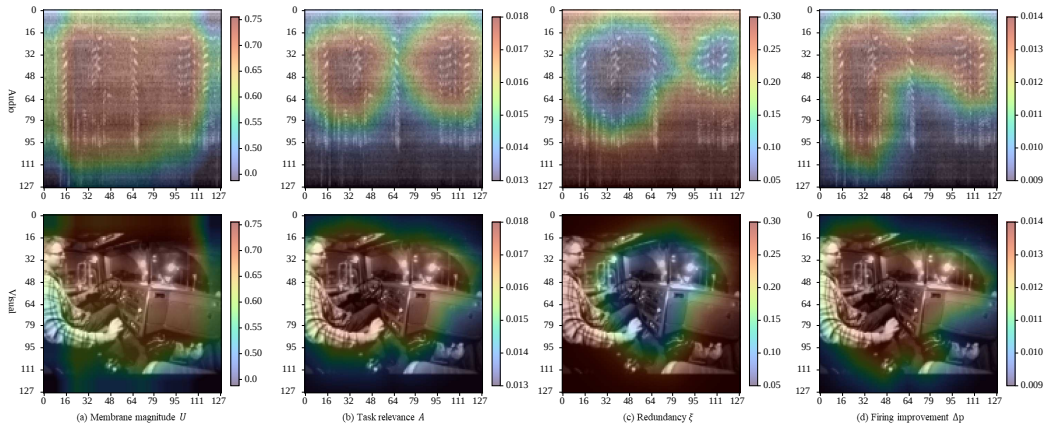

Figure 3: Visualization of variables in ST-MAT on AVE

Table 2: Ablation study on AVE

Method	Acc. (%)	Δ
ST-MAT	**78.47**	–
w/o TRSSE	75.25	-3.22
w/o SRB	75.50	-2.97
w/o TRSSE & SRB	74.01	-4.46
Linear utility	73.27	-5.20
w/o SBTA	72.28	-6.19
Firing-controlled (matched ST-MAT)	71.29	-7.18

form allows one term to compensate for a poor value of the other, whereas the multiplicative form penalizes cases in which either semantic retention or redundancy rejection becomes excessively low. Removing SBTA, shown in the sixth row, reduces accuracy to 72.28%, a 6.19% decrease from ST-MAT. Direct threshold replacement changes the spike distribution too quickly and disrupts the stability of the statistics used to select the next threshold. We further use a loss term to constrain an SNN with the same network architecture to maintain the same firing rate as ST-MAT. The final row achieves only 71.29%, which is 7.18% below ST-MAT. This shows that matching firing rate does not reproduce the same learning effect and supports our motivation that spike quantity alone is insufficient. ST-MAT also improves unimodal SNNs, as shown in Appendix Table A4. Additional results on threshold learning at different stages and different cross-attention placements are provided in Appendix E.

6 Conclusion

The rich spatiotemporal dynamics of spiking neural networks make them well suited to audio-visual fusion. Existing AV-SNN studies typically assume that continuous internal neuronal representations have already been appropriately discretized into spikes. We argue that the number of spikes produced by a modality does not determine AV-SNN performance. Instead, it is necessary to distinguish which internal neuronal representations carry task-discriminative information and which merely repeat existing spatiotemporal information. To address this problem, we propose Spatio-Temporal Modality-Adaptive Thresholding (ST-MAT). TRSSE selects spatiotemporal locations with high task relevance, while SRB distinguishes repeatedly expressed redundant representations. SBTA then determines which continuous responses should be converted into spikes through slow and bounded threshold learning. We validate ST-MAT on AVE, CREMA-D, and UrbanSound8K-AV. ST-MAT consistently outperforms both fixed-threshold and ordinary learnable-threshold variants and shows clear advantages over competing SNN methods, reaching 78.47% on AVE and achieving the state of the art among the reported SNN methods. Visualization and ablation results show that ST-MAT converts high-task-value internal neuronal representations into spikes while reducing redundancy, thereby substantially improving AV-SNN fusion. ST-MAT introduces no additional inference overhead, making it promising for deployment on neuromorphic hardware for multisensory classification tasks.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and manuscript presentation, including improving grammar, readability, conciseness, terminology consistency, and the clarity of technical descriptions. They were also used to assist in identifying potentially relevant literature, which was subsequently checked and verified by the authors. The research motivation, problem formulation, methodological design of ST-MAT, theoretical analysis, experimental setup, implementation, result interpretation, and scientific conclusions were independently determined by the authors. All AI-assisted revisions were carefully reviewed by the authors, who take full responsibility for the accuracy, integrity, and final content of this work.

REPRODUCIBILITY STATEMENT

Appendix C reports preprocessing, optimization, calibration, and deployment-computation details, while Appendices D–F contain the supplementary analyses referenced from the main experiments. The anonymous code is available at `https://anonymous.4open.science/r/ST-MAET-3F33`.

REFERENCES

Azouz, R. and Gray, C. M. Dynamic spike threshold reveals a mechanism for synaptic coincidence detection in cortical neurons in vivo. *Proceedings of the National Academy of Sciences*, 97(14):8110–8115, 2000.

Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., and Maass, W. Long short-term memory and learning-to-learn in networks of spiking neurons. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature Neuroscience*, 2:515–520, 1999.

Fan, Y., Xu, W., Wang, H., Wang, J., and Guo, S. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038, 2023.

Fang, W., Yu, Z., Chen, Y., Masquelier, T., Huang, T., and Tian, Y. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2661–2671, 2021.

Fontaine, B., Peña, J. L., and Brette, R. Spike-threshold adaptation predicted by membrane potential dynamics in vivo. *PLoS Computational Biology*, 10(4):e1003560, 2014.

Grubb, M. S. and Burrone, J. Activity-dependent relocation of the axon initial segment fine-tunes neuronal excitability. *Nature*, 465:1070–1074, 2010.

Guo, L., Gao, Z., Qu, J., Zheng, S., Jiang, R., Lu, Y., and Qiao, H. Transformer-based spiking neural networks for multimodal audio-visual classification. *IEEE Transactions on Cognitive and Developmental Systems*, 16(3):1077–1086, 2024.

He, X., Zhao, D., Dong, Y., Shen, G., Yang, X., and Zeng, Y. Enhancing audio-visual spiking neural networks through semantic-alignment and cross-modal residual learning. *arXiv preprint arXiv:2502.12488*, 2025.

He, X., Zhao, D., Li, Y., Kong, Q., Yang, X., and Zeng, Y. Incorporating brain-inspired mechanisms for multimodal learning in artificial intelligence. *Science Advances*, 12(23):eady8751, 2026.

Jiang, R., Han, J., Xue, Y., Wang, P., and Tang, H. CMCI: A robust multimodal fusion method for spiking neural networks. In *Neural Information Processing (ICONIP 2023)*, pp. 159–171, Springer, 2023.

Li, H., Li, X., Hu, P., Lei, Y., Li, C., and Zhou, Y. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Liu, Q., Xing, D., Feng, L., Tang, H., and Pan, G. Event-based multimodal spiking neural network with attention mechanism. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8922–8926, 2022.

Liu, X., Xia, N., Zhou, J., Li, Z., and Guo, D. Towards energy-efficient audio-visual classification via multimodal interactive spiking neural network. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(5):144:1–144:24, 2025.

Ma, N., Qu, B., Wang, W., Wang, F., and Zhang, X. SCAF-Net: A spiking cross-modal attention fusion network for multimodal emotion recognition. *Biomedical Signal Processing and Control*, 118:109684, 2026.

Ma, N., Zhang, X., Qu, B., and Wang, W. EAVS-TF: A bio-inspired spiking neural network for energy-efficient multimodal emotion recognition. *Current Radiopharmaceuticals*, 19(3):100041, 2026.

Neftci, E. O., Mostafa, H., and Zenke, F. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Oikonomou, K. M., Kansizoglou, I., and Gasteratos, A. Spiking neural network feature discrimination boosts modality fusion. *IEEE Transactions on Cognitive and Developmental Systems*, 18(4):943–955, 2026.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Rohe, T. and Noppeney, U. Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Current Biology*, 26(4):509–514, 2016.

Saleh, M. and Ahmadi, Z. Cross-modal binary attention: An energy-efficient fusion framework for audio-visual learning. *arXiv preprint arXiv:2602.00701*, 2026.

Senkowski, D. and Engel, A. K. Multi-timescale neural dynamics for multisensory integration. *Nature Reviews Neuroscience*, 25(9):625–642, 2024.

Shen, J., Xie, Y., Xu, Q., Pan, G., Tang, H., and Chen, B. Spiking neural networks with temporal attention-guided adaptive fusion for imbalanced multi-modal learning. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, 2025. doi:10.1145/3746027.3755622.

Stein, B. E. and Stanford, T. R. Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9:255–266, 2008.

Tian, Y., Shi, J., Li, B., Duan, Z., and Xu, C. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 247–263, 2018.

Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J. T., and Peng, X. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769, 2023.

Zhong, J., Le, V. T. D., Pham, H. L., and Nakashima, Y. Spikceiver: Compact audio-visual spiking transformer with depthwise splitting and bottleneck attention. In *Proceedings of the IEEE Symposium on Low-Power and High-Speed Chips and Systems (COOL CHIPS)*, pp. 1–6, 2026.

A Theoretical Analysis and Proofs

This appendix provides the proofs referenced in Sec. 4. The results establish the boundedness of s_{ST}, the consistency between global and local redundancy, the effect of support on threshold preference, the non-equivalence between firing rate and effective support, and the stability properties of the geometric utility and bounded update.

Proposition 1 (Bounded effective support)**.** *Assume $w_i \geq 0$, $\sum_i w_i = 1$, $C_{ii} = 1$ on nonzero-mass sites, and $0 \leq C_{ij} \leq 1$ over $M = T \cdot N_l$ sites. Then $0 \leq s_{\mathrm{ST}} \leq 1$.*

Proof. Since $C_{ij}^2 \leq 1$, $\mathcal{R}_{\mathrm{ST}} \leq \sum_{ij} w_i w_j = 1$. Diagonal terms give $\mathcal{R}_{\mathrm{ST}} \geq \sum_i w_i^2$, and Cauchy–Schwarz yields $\sum_i w_i^2 \geq (\sum_i w_i)^2/M = 1/M$. Hence $1/M \leq \mathcal{R}_{\mathrm{ST}} \leq 1$, $1 \leq d_{\mathrm{ST}} \leq M$, and therefore $0 \leq s_{\mathrm{ST}} \leq 1$. □

Proposition 2 (Global–local redundancy consistency)**.** *With $\xi_i = \sum_j w_j C_{ij}^2$, the global redundancy equals the task-weighted expectation of local redundancy: $\mathcal{R}_{\mathrm{ST}} = \sum_i w_i \xi_i$.*

Proof. Direct substitution gives $\sum_i w_i \xi_i = \sum_i w_i \sum_j w_j C_{ij}^2 = \sum_{ij} w_i w_j C_{ij}^2 = \mathcal{R}_{\mathrm{ST}}$. □

Proposition 3 (Support-dependent threshold preference)**.** *For two candidate thresholds $\theta_1 < \theta_2$ with positive semantic and redundancy terms, the relative log-utility of θ_1 is nondecreasing in α and therefore in s_{ST}.*

Proof. Define $D(\alpha) = \log J_\alpha(\theta_1) - \log J_\alpha(\theta_2)$. Because the lower threshold passes a superset of states, $R_{\mathrm{sem}}(\theta_1) \geq R_{\mathrm{sem}}(\theta_2)$ and $R_{\mathrm{red}}(\theta_1) \leq R_{\mathrm{red}}(\theta_2)$. Therefore

$$D'(\alpha) = \log \frac{R_{\mathrm{sem}}(\theta_1)}{R_{\mathrm{sem}}(\theta_2)} - \log \frac{R_{\mathrm{red}}(\theta_1)}{R_{\mathrm{red}}(\theta_2)} \geq 0.$$

Since $d\alpha/ds_{\mathrm{ST}} = 1/2 > 0$, increasing support increases the relative preference for the lower threshold, all else fixed. This is a pairwise preference result and does not imply a universal ordering between audio and visual thresholds because their continuous-response and relevance distributions differ. □

Proposition 4 (Firing rate is not equivalent to effective support)**.** *There is no general function $s_{\mathrm{ST}} = f(r)$. Equal spike counts can correspond to different effective supports.*

Proof. Consider two sites with equal task weights. If $C_{12} = 1$, then $\mathcal{R}_{\mathrm{ST}} = 1$ and $d_{\mathrm{ST}} = 1$. If $C_{12} = 0$ and the descriptors are orthogonal, then $\mathcal{R}_{\mathrm{ST}} = 1/2$ and $d_{\mathrm{ST}} = 2$. The two representations may emit the same total number of spikes, hence the same firing rate, while their effective supports differ. □

Proposition 5 (Non-trivial geometric utility)**.** *For $0 < \alpha < 1$ and nondegenerate semantic and redundancy distributions, neither the all-pass nor the all-reject threshold is optimal whenever an interior threshold yields both $R_{\mathrm{sem}} > 0$ and $R_{\mathrm{red}} > 0$.*

Proof. An all-pass threshold gives $R_{\mathrm{sem}} = 1, R_{\mathrm{red}} = 0$, hence $J = 0$. An all-reject threshold gives $R_{\mathrm{sem}} = 0, R_{\mathrm{red}} = 1$, again $J = 0$. Any interior threshold with both factors positive has $J > 0$ and strictly dominates the endpoints. □

Proposition 6 (Contraction of the slow threshold update)**.** *If clipping and projection are inactive, θ^* is fixed over one calibration interval, and $0 < \eta_\theta < 1$, then the update in Eq. 6 contracts the threshold error geometrically.*

Proof. Eq. 6 reduces to $\theta^{q+1} = \theta^q + \eta_\theta(\theta^* - \theta^q)$, so $\theta^{q+1} - \theta^* = (1 - \eta_\theta)(\theta^q - \theta^*)$. Since $|1 - \eta_\theta| < 1$, the error contracts geometrically. Clipping further bounds every update by Δ_θ, while projection enforces the predefined safe interval. □

B Algorithm

Training loop. For each minibatch, perform the ordinary AV-SNN forward pass with current stage-wise thresholds, execute QK/QK/QKV fusion, compute the task loss, and update model parameters with surrogate-gradient backpropagation. At designated training epochs, collect continuous pre-spike representations from a calibration subset, compute modality-specific margins and relevance tensors, form w, Z, C, s_{ST}, and ξ, construct p^+ and p^-, sort candidate response values, evaluate semantic-retention and redundancy-rejection curves, select θ^*, and apply the bounded update in Eq. 6. At inference, only the fixed learned thresholds remain.

C Implementation Details

This appendix gives the detailed experimental configuration used for the reported models, including preprocessing, architecture dimensions, optimization, threshold updating, and initialization.

Data preprocessing. Raw audio–visual samples are first organized by class label. For every sample, audio is converted to mono and resampled to 16 kHz, then uniformly divided into six temporal segments. A 128-bin Mel power spectrogram is extracted from each segment with hop length 512, converted to dB, normalized by per-sample min–max scaling, and resized to 128×128. Six

visual frames are sampled at the corresponding temporal positions, converted to grayscale, bilinearly resized to 128×128, and normalized to $[0, 1]$. UrbanSound8K-AV uses the established audio–visual pairing to retrieve its visual sequences. Both modalities are stored as float32 NPY arrays of shape $[N, 6, 1, 128, 128]$, and labels are converted to zero-based integers. During training, the two modalities use the same sample permutation and synchronized CutMix with $\alpha = 0.3$; no random augmentation is applied for validation or testing.

Network configuration. The three datasets use separate audio and visual SPS branches with one, one, and two blocks across the three stages. Stages 1 and 2 keep the two modalities independent: each branch is processed by its own QK spiking-attention block and spiking MLP, with no cross-modal connection. Cross-modal interaction is introduced only in Stage 3 through the bidirectional QKV spiking-attention block, consistent with Fig. 2. Every stage uses eight attention heads and an MLP expansion ratio of 4. The base dimension d is 96, 192, and 96 for AVE, CREMA-D, and UrbanSound8K-AV, respectively, giving stage widths $(d, 2d, 4d)$. Final audio and visual features are spatially averaged, the two modalities are averaged by Add fusion, and a linear classifier predicts 28, 6, and 10 classes. The semantic-alignment projection dimension is 128. Table A1 lists the spatial resolution, channel width, per-head dimension, and MLP width at each stage.

Table A1: Dual-branch network structure and dimensions for the three datasets. Entries are reported per modality.

Module	AVE	CREMA-D	UrbanSound8K-AV
Input	$(128\times128, 1) \times 6$	$(128\times128, 1) \times 6$	$(128\times128, 1) \times 6$
SPS stem	$(64\times64, 48) \rightarrow (32\times32, 96)$	$(64\times64, 96) \rightarrow (32\times32, 192)$	$(64\times64, 48) \rightarrow (32\times32, 96)$
Stage 1 QK $\times 1$	$(32\times32, 96)$ 8 heads $\times$ 12; MLP 384	$(32\times32, 192)$ 8 heads $\times$ 24; MLP 768	$(32\times32, 96)$ 8 heads $\times$ 12; MLP 384
Stage 2 QK $\times 1$	$(16\times16, 192)$ 8 heads $\times$ 24; MLP 768	$(16\times16, 384)$ 8 heads $\times$ 48; MLP 1536	$(16\times16, 192)$ 8 heads $\times$ 24; MLP 768
Stage 3 QKV $\times 2$	$(8\times8, 384)$ 8 heads $\times$ 48; MLP 1536	$(8\times8, 768)$ 8 heads $\times$ 96; MLP 3072	$(8\times8, 384)$ 8 heads $\times$ 48; MLP 1536
Pooling/output	Add $384 \rightarrow 28$	Add $768 \rightarrow 6$	Add $384 \rightarrow 10$

Training, threshold updating, and initialization. All three datasets are trained for 150 epochs with four-GPU distributed training. The per-GPU batch size is 32 for AVE and UrbanSound8K-AV and 16 for CREMA-D. We use AdamW with a command-line learning rate of 0.008, weight decay 10^{-4}, cosine scheduling, five warm-up epochs, and random seed 42. LIF neurons use an initial threshold of 0.5 and a membrane time constant of 2.0. The upper bounds of the stage-wise residual coefficients for Stages 1–3 are 0.2, 0.4, and 0.6, respectively.

The ST-MAT configuration models task relevance and redundancy at all three stages, uses the geometric objective, and applies trust-region threshold updates. Threshold statistics are re-estimated every five epochs from 64 class-stratified samples with estimation batch size 2. The threshold update coefficient is 0.15, the maximum single update is 0.5% of the initial threshold, and the trust-region boundary is 5% of the initial threshold. Convolutional and linear layers use the PyTorch default initialization, BatchNorm scale and bias are initialized to 1 and 0, and no pretrained weights are loaded.

Computation overhead. We compare ST-MAT with the ordinary learnable-threshold and fixed-threshold variants on the same AVE backbone. Table A1b summarizes the measured training computation, inference computation, and parameter count. ST-MAT requires 97.79 G FLOPs during training, only 0.13 G more than ST-MAT with learnable threshold and ST-MAT with fixed threshold, both of which require 97.66 G FLOPs. Thus, ST-MAT introduces only a very small amount of additional computation during training. At inference, all three methods require 32.61 G FLOPs, showing that ST-MAT introduces no additional inference computation. Their parameter counts are also identical at 16.35 M, indicating that ST-MAT does not change the model parameter count.

The measured difference can be understood from the computation structure. Let $F_{\mathrm{net}}^{\mathrm{train}}$ denote the FLOPs of the ordinary backbone forward/backward pass for one training minibatch. The fixed-threshold variant does not update neuronal thresholds, so its training cost is

$$F_{\mathrm{FT}}^{\mathrm{train}} = F_{\mathrm{net}}^{\mathrm{train}}. \tag{7}$$

Table A1b: Computation overhead on the matched AVE backbone.

Method	Training FLOPs (G)	Inference FLOPs (G)	Params (M)
ST-MAT	97.79	32.61	16.35
ST-MAT with learnable threshold	97.66	32.61	16.35
ST-MAT with fixed threshold	97.66	32.61	16.35

For the ordinary learnable-threshold variant, threshold learning adds only elementwise threshold-gradient operations, denoted by F_θ, giving

$$F_{\mathrm{LT}}^{\mathrm{train}} = F_{\mathrm{net}}^{\mathrm{train}} + F_\theta, \qquad F_\theta = \mathcal{O}\left(\sum_l BTN_l D_l\right), \tag{8}$$

where B is the minibatch size and T, N_l, and D_l are the temporal length, number of spatial/token locations, and channel dimension at stage l.

ST-MAT does not add a new forward branch to every training minibatch. Its extra computation appears only at periodic threshold-update steps. For one modality-stage pair, define $R_l = TN_l$ and let B_c be the batch size used for estimating the threshold statistics. Constructing the relevance-weighted descriptors costs $\mathcal{O}(B_c R_l D_l)$, while the pairwise similarity matrix used by TRSSE and SRB costs

$$F_{\mathrm{sim},l} = \mathcal{O}\big(B_c R_l^2 D_l\big)\,. \tag{9}$$

If $K_l = R_l D_l$ observed responses are used as threshold candidates, sorting them and evaluating the cumulative semantic-retention and redundancy-rejection terms costs approximately $\mathcal{O}(B_c K_l \log K_l)$; the bounded threshold update itself is constant-time per modality-stage pair. Let Q denote the number of ordinary training minibatches between two threshold-update events. Averaged over training, the ST-MAT cost is therefore

$$F_{\mathrm{ST}}^{\mathrm{train}} \approx F_{\mathrm{net}}^{\mathrm{train}} + \frac{1}{Q}\sum_{m,l}\left(F_{\mathrm{rel},m,l} + F_{\mathrm{sim},m,l} + F_{\mathrm{sort},m,l}\right). \tag{10}$$

Because these operations are performed only periodically on a small subset, the amortized second term is small relative to the backbone forward/backward computation, which is consistent with the measured 97.79 G versus 97.66 G training FLOPs in Table A1b.

At inference, task-relevance estimation, redundancy analysis, threshold search, and threshold updates are removed, and only the learned threshold values remain. Therefore, all three variants execute the same deployed backbone computation,

$$F_{\mathrm{ST}}^{\mathrm{infer}} = F_{\mathrm{LT}}^{\mathrm{infer}} = F_{\mathrm{FT}}^{\mathrm{infer}} = F_{\mathrm{net}}^{\mathrm{infer}}, \tag{11}$$

which agrees with the identical 32.61 G inference FLOPs in Table A1b. ST-MAT also introduces no additional model parameters, so

$$P_{\mathrm{ST}} = P_{\mathrm{LT}} = P_{\mathrm{FT}} = 16.35\ \mathrm{M}. \tag{12}$$

Thus, ST-MAT adds only a small training-time computation, introduces no additional inference computation, and leaves the model parameter count unchanged.

D Firing Rate

Table A2 reports recognition accuracy together with the stage-wise audio and visual firing rates of the final ST-MAT models. On AVE, ST-MAT reaches 78.47% accuracy while the audio and visual firing rates are 15.9/17.6%, 16.0/16.6%, and 19.0/19.4% across Stages 1–3, corresponding to signed differences of $-1.7/-0.6/-0.4$ percentage points. On CREMA-D, ST-MAT reaches 79.44% accuracy with firing-rate differences of $-1.2/-1.7/-0.7$ percentage points across the three stages. On UrbanSound8K-AV, ST-MAT reaches 99.54% accuracy, while the corresponding firing-rate differences are $-1.4/-2.0/-4.6$ percentage points. These results show that strong recognition performance does not require a fixed audio–visual firing-rate balance, and that the firing pattern varies across datasets and stages.

Table A2: Recognition accuracy and stage-wise firing rates of the final ST-MAT models. Firing-rate changes are signed percentage-point differences with $\Delta r = r_A - r_V$.

Dataset	Stage	Acc. (%)	r_A (%)	r_V (%)	Δr
AVE	1	78.47	15.9	17.6	-1.7
	2		16.0	16.6	-0.6
	3		19.0	19.4	-0.4
CREMA-D	1	79.44	17.5	18.7	-1.2
	2		16.8	18.5	-1.7
	3		17.1	17.8	-0.7
UrbanSound8K-AV	1	99.54	17.6	19.0	-1.4
	2		18.1	20.1	-2.0
	3		21.0	25.6	-4.6

Table A2b: Dataset-wise firing-rate validation and recognition accuracy. ST-MAT firing rates are averages over Stages 1–3 and are reported as percentages.

Dataset	Method	Audio firing (%)	Visual firing (%)	Acc. (%)
AVE	ST-MAT with fixed threshold	9.14	11.64	73.27
	Firing-controlled (matched ST-MAT)	18.29	18.91	71.29
	ST-MAT	16.97	17.87	78.47
CREMA-D	ST-MAT with fixed threshold	12.19	7.43	76.21
	ST-MAT	17.13	18.33	79.44
UrbanSound8K-AV	ST-MAT with fixed threshold	11.99	14.23	98.40
	ST-MAT	18.90	21.57	99.54

Table A2b further compares dataset-wise firing rates and recognition accuracy. For each ST-MAT row, the audio and visual firing rates are averaged over Stages 1–3 so that each dataset is summarized by one operating point.

The firing-controlled AVE run is particularly informative. Its audio and visual firing rates, 18.29% and 18.91%, are close to the ST-MAT operating rates of 16.97% and 17.87%, yet its accuracy is only 71.29%, compared with 78.47% for ST-MAT. The fixed-threshold variant has lower firing rates of 9.14% and 11.64% and reaches 73.27% accuracy. Thus, neither simply increasing firing activity nor matching the firing density of ST-MAT reproduces its recognition performance. The same tendency is observed on CREMA-D and UrbanSound8K-AV, where the CREMA-D fixed-threshold variant has audio and visual firing rates of 12.19% and 7.43% with 76.21% accuracy, while ST-MAT reaches 79.44%; on UrbanSound8K-AV, accuracy improves from 98.40% to 99.54%. These results indicate that recognition accuracy is not determined by spike quantity alone.

E Cross-Attention Placement and Stage-Wise Threshold Ablations

Table A3 reorganizes the AVE ablations by jointly varying where cross-modal attention is introduced and where ST-MAT threshold adaptation is enabled. When cross-modal attention is used only in Stage 3, full ST-MAT obtains 78.47% test accuracy (76.49% train), while adapting Stage 3 alone also reaches 78.47% (76.24% train). In comparison, adapting only Stage 1 or Stage 2 gives 75.00% and 76.98%, respectively. For the adopted backbone, the strongest threshold-adaptation effect is concentrated at the late semantic stage associated with the designated cross-modal interaction.

Moving cross-modal attention to earlier stages changes this behavior. With cross-modal attention in Stages 2–3, enabling adaptation in all three stages gives 74.50% test accuracy, while Stage-1-only, Stage-2-only, and Stage-3-only adaptation give 75.25%, 72.03%, and 72.03%, respectively. When cross-modal attention is present in all three stages, all-stage adaptation gives 73.76%, adapting Stages 2–3 gives 72.28%, and adapting Stage 3 alone gives 75.74%. Across these configurations, the best test result is obtained when cross-modal interaction is restricted to Stage 3; within that configuration, Stage-3-only adaptation matches full ST-MAT. These results favor keeping the early stages modality-specific and concentrating threshold adaptation near the late cross-modal interaction stage.

Table A3: Cross-attention placement and stage-wise threshold-adaptation ablations on AVE. The Group column specifies the stages containing cross-modal attention; the Stage 1–3 columns indicate where ST-MAT threshold adaptation is enabled. Test and training accuracies are reported directly from the corresponding experimental runs.

Cross-attention group	Adaptation setting	Stage 1	Stage 2	Stage 3	Test (%)	Train (%)
Stage 3 only	ST-MAT	✓	✓	✓	78.47	76.49
	Stage 1 only	✓	✗	✗	75.00	75.00
	Stage 2 only	✗	✓	✗	76.98	75.25
	Stage 3 only	✗	✗	✓	78.47	76.24
Stages 2–3	Stages 1–3	✓	✓	✓	74.50	74.75
	Stage 1 only	✓	✗	✗	75.25	74.51
	Stage 2 only	✗	✓	✗	72.03	74.26
	Stage 3 only	✗	✗	✓	72.03	74.26
Stages 1–3	Stages 1–3	✓	✓	✓	73.76	74.75
	Stages 2–3	✗	✓	✓	72.28	74.26
	Stage 3 only	✗	✗	✓	75.74	74.26

Table A4: Unimodal comparison on AVE with the same branch architecture.

Method	Audio (%)	Visual (%)	Avg. gain
ST-MAT with fixed threshold	69.65	35.07	–
ST-MAT	**72.03**	**38.86**	**+3.09**

F Unimodal Comparison

A direct unimodal test asks whether ST-MAT improves the representation of each branch even before audiovisual fusion. Table A4 shows that on AVE, audio-only accuracy increases from 69.65% to 72.03% (+2.38%), while visual-only accuracy increases from 35.07% to 38.86% (+3.79%). This is consistent with ST-MAT acting on representation formation rather than only on fusion. Each branch becomes more discriminative when its continuous internal responses are converted to spikes according to task relevance and redundancy. The ST-MAT with fixed threshold diagnostic also exhibits strongly task-dependent natural activity (e.g., hard firing rates of 30.6% for AVE audio and 1.5% for AVE visual, versus 30.9% for CREMA-D visual), which motivates avoiding fixed modality-level firing-rate priors.

The cross-attention-placement and stage-wise threshold controls are consolidated in Table A3. Across the completed variants, the strongest test result is obtained when early stages remain modality-specific and cross-modal interaction is introduced only at Stage 3.

G Biological Motivation for ST-MAT

ST-MAT is inspired by biological fast neuronal responses and slower control of excitability. Cortical spike thresholds are not immutable constants; they vary with recent membrane trajectories and depolarization dynamics (Azouz and Gray, 2000; Fontaine et al., 2014). This suggests that the boundary between a neuron's continuous internal activity and its emitted spike representation can depend on the current operating regime rather than being fixed globally. ST-MAT adopts that limited principle by examining the continuous pre-spike representation $U_{m,l}$ before discretization and using task relevance to decide where the threshold should move.

A second observation is that intrinsic excitability changes on a slower timescale than individual spikes and fast synaptic responses (Desai et al., 1999; Grubb and Burrone, 2010). This motivates the separation between ordinary minibatch learning and periodic threshold calibration. The preferred threshold is not applied abruptly. SBTA moves toward it with a bounded step and projection, reflecting the distinction between fast response dynamics and slower excitability regulation.

Neural responses depend on when and where signals occur and on their behavioral relevance (Stein and Stanford, 2008; Rohe and Noppeney, 2016; Senkowski and Engel, 2024). For ST-MAT, this motivates looking beyond global firing density. TRSSE measures how task-relevant information is distributed over time and token position after correlated repetitions are discounted. SRB then combines that global support with local redundancy and task relevance so that a response is not

suppressed merely because it is correlated with another one. The resulting threshold rule preserves useful activity while reducing repeated low-value responses, and SBTA makes that rule evolve slowly as the network representation changes.