Mutual information-improved audio-visual segmentation using spiking neural networks

Guoyu Zuo[1], Zhenyu Yang[1,2], Yitao Qin[3], Erjun Xiao[1], and Tielin Zhang[2⋆]

[1] School of Information Science and Technology, Beijing University of Technology,and Beijing Key Laboratory of Computing Intelligence and Intelligent Systems, China
[2] Center for Excellence in Brain Science and Intelligence Technology, State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Chinese Academy of Sciences, China
[3] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, China
zhangtielin@ion.ac.cn

Abstract. Audio-visual segmentation (AVS) segments sounding objects by aligning auditory cues with visual regions. Existing AVS systems are dominated by artificial neural networks, while SNN-based dense multimodal segmentation remains underexplored. This paper introduces MI-SNN, an SNN AVS framework that improves an AVSegFormer-style SNN baseline through mutual-information-guided fusion. The baseline converts the AVSegFormer audio-visual pipeline into spike-based representations while keeping its encoders, query interaction, and decoder. It verifies feasibility but suffers a large degradation because spike discretization weakens fine-grained cross-modal correspondence. MI-SNN estimates dependency between audio and spatial visual features and enhances sound-consistent regions. On the S4 and MS3 subsets of AVS-Bench, the SNN baseline achieves 62.50 mIoU on S4, whereas MI-SNN improves it to 76.97 mIoU, showing that information-theoretic alignment is effective for this SNN-based AVS setting.

Keywords: Audio-visual segmentation · Mutual information · Spiking neural networks · Multimodal learning

1 Introduction

Humans localize sounding objects by jointly interpreting what is seen and what is heard. Audio-visual segmentation (AVS) formalizes this ability as a dense prediction task: given video frames and synchronized audio, a model predicts pixel-level masks for objects that produce sound [40]. Compared with image segmentation, AVS requires a model to solve two coupled problems. First, it must preserve spatial details that are necessary for accurate masks. Second, it must determine which visual regions are semantically related to the audio signal. This second requirement makes AVS a representative multimodal learning problem,

⋆ Corresponding author.

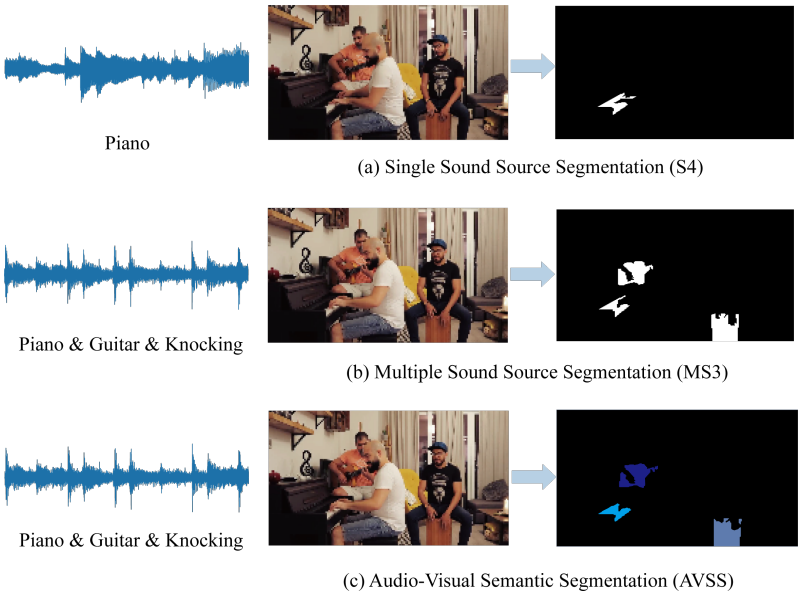

Fig. 1. Illustration of AVSBench task settings. From top to bottom, S4 predicts a binary mask for one sounding source, MS3 predicts binary masks for multiple sounding sources, and AVSS predicts semantic masks with category labels.

because the visual foreground cannot be reliably inferred from the image stream alone when multiple objects or distractors are present.

AVSBench [40] defines three representative settings, as shown in Fig. 1: S4 for single-source segmentation, MS3 for multiple simultaneous sources, and AVSS for semantic categories. Related tasks include correspondence learning [1], sound-source localization [29,4,28,36], and dense audio-visual event localization [10]. Recent AVS methods strengthen cross-modal interaction: AVSegFormer [9] uses audio-visual mixers and audio-driven queries; TransAVS [20] improves query disentanglement and self-supervised alignment; AVESFormer [35] enhances transformer design; and contrastive latent diffusion explores generative AVS modeling [23]. These ANN-based methods show that accurate AVS depends on fine-grained audio-visual correspondence, while the potential of SNNs for dense multimodal segmentation remains unclear.

SNNs encode information with spikes and membrane-potential dynamics. Recent studies have explored spike-based models for temporal and visual perception, including semantic segmentation [17,12], motion segmentation [26], ANN-to-SNN conversion [8], recurrent learning [16], spiking speech recognition [32], and spike-driven transformer architectures [38]. Nevertheless, AVS poses a different challenge: the model must preserve dense spatial semantics while using audio to select sound-related pixels. A direct SNN conversion is therefore insufficient. Spike discretization can reduce the precision of hidden representations, and this reduction is especially harmful for AVS because the task relies on subtle audio-visual dependencies rather than single-modal recognition alone.

We study SNNs for AVS from a baseline-to-method perspective. We first build an AVSegFormer-SNN baseline by converting the AVSegFormer-style pipeline into a spike-based version without an explicit mutual-information objective. This baseline produces valid masks but remains much weaker than strong ANN-based AVS models. We then propose MI-SNN, which keeps the AVSegFormer-SNN formulation and adds mutual-information-guided fusion. Inspired by MINE [2], MI-SNN estimates dependency between audio representations and spatial visual features, and uses the resulting dependency map to enhance sound-related regions before spiking mask decoding.

Our contributions are threefold. First, we present an early SNN framework for AVSBench-style dense audio-visual segmentation. Second, we establish an AVSegFormer-SNN baseline and show that direct spike conversion causes a severe performance drop. Third, we introduce MI-guided fusion and demonstrate substantial gains over the SNN baseline on the S4 and MS3 subsets of AVSBench.

2 Related Work

2.1 Audio-visual segmentation

AVS locates sounding objects at the pixel level by combining audio and visual information [40]. Early baselines showed that temporal interaction and audio-conditioned reasoning help separate sounding objects from silent distractors. Later methods strengthened interaction with audio-visual mixers and query decoding [9,20], instance association [6], bilateral relations [37], selective learning [18], consistency-queried transformers [21], class-conditional prompting [7], progressive semantic training [22], textual semantics [34], bias mitigation [30], foundation-model localization [24,3], and stronger audio or vision-centric cues [5,15]. These works show that AVS requires audio semantics to be aligned with localized visual evidence.

2.2 Mutual information for multimodal alignment

Mutual information measures statistical dependence between variables and is therefore well suited for multimodal alignment. MINE [2] estimates mutual information with a neural discriminator and the Donsker–Varadhan lower bound, making it applicable to high-dimensional representations. Variational bounds, MI analysis, and MI maximization have also been used to improve multimodal representation learning [27,14,31,11,25,39,19]. In AVS, the relevant dependency is not merely between the input audio and the final mask. Instead, the model must determine which spatial visual tokens are related to the audio signal. We therefore use mutual information as a hidden-state alignment objective between global audio features and local visual features. This formulation directly encourages the network to assign high dependency to visual regions that correspond to the sound source.

2.3 Spiking neural networks for dense multimodal prediction

SNNs process information with spike events and membrane states. Although spiking models have been applied to several perception tasks, multimodal dense prediction remains underexplored. Spike firing approximation (SFA) [38] provides a practical training strategy by representing activations as integer spike counts during training and expanding them into binary spike trains during inference. This makes it possible to build deeper spiking transformer modules. However, AVS requires more than spike-based feature extraction. An AVSegFormer-SNN baseline may preserve coarse object responses but fail to maintain the fine-grained dependency between audio cues and visual pixels. Our work addresses this gap by combining SNN representations with an explicit mutual-information alignment objective.

3 Method

3.1 Problem formulation and model overview

Given a video clip $V = \{v_t\}_{t=1}^{T}$ and synchronized audio segments $A = \{a_t\}_{t=1}^{T}$, AVS predicts a mask $\hat{Y}_t$ for each frame with ground truth Y_t. The prediction should identify objects that are both visible and acoustically active. Fig. 2 shows the architecture of MI-SNN. The model consists of a visual encoder, an audio encoder, an MI-guided fusion module, a spiking transformer, and a segmentation decoder. We evaluate two variants. The first is an AVSegFormer-SNN baseline that converts the AVSegFormer audio-visual segmentation pipeline into an SNN and removes the MI-guided objective. The second is MI-SNN, our method, which keeps this SNN backbone and adds MI-guided fusion to strengthen audio-visual alignment.

3.2 Audio and visual feature extraction

The visual stream uses PVTv2 [33] to extract hierarchical frame features $\mathcal{F}_v \in \mathbb{R}^{B_c \times T \times C \times H \times W}$ from B_c clips and T sampled frames. The audio stream converts synchronized audio segments into log-mel spectrograms and uses VGGish [13] to obtain $\mathcal{F}_a \in \mathbb{R}^{B_c \times T \times C}$. We fold clip and time axes into $N = B_c T$ frame-audio pairs, so the equations use $F_v \in \mathbb{R}^{N \times C \times H \times W}$ and $F_a \in \mathbb{R}^{N \times C}$. Each audio feature acts as a global semantic query for its synchronized frame.

To build the SNN versions, dense activations in the AVSegFormer-style audio-visual pipeline are replaced by spike-based hidden representations. The AVSegFormer-SNN baseline performs this replacement without adding an explicit cross-modal dependency objective. Therefore, it tests whether spike-based representations alone are sufficient for AVS. MI-SNN uses the same spiking representation but introduces the MI-guided fusion described below.

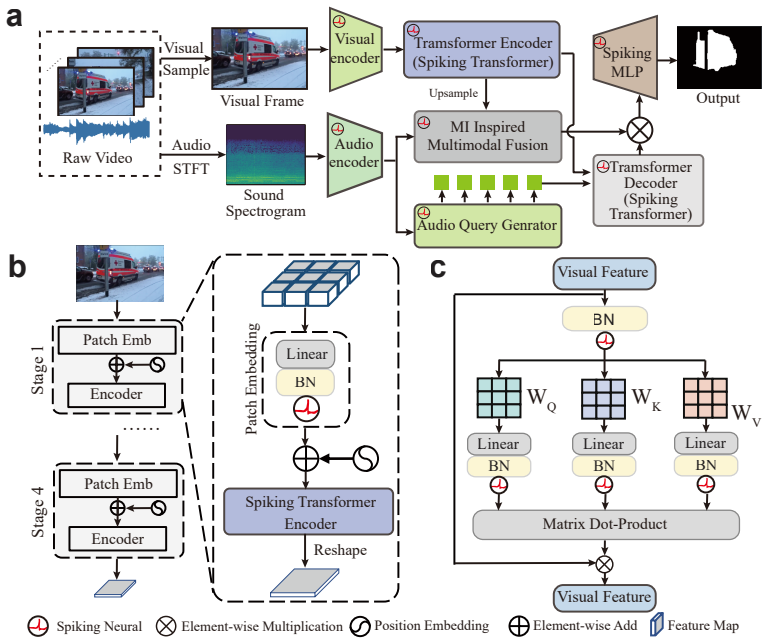

Fig. 2. Architecture of MI-SNN. The AVSegFormer-SNN baseline removes the MI-guided fusion objective, whereas MI-SNN estimates audio-visual dependency and enhances sound-related visual regions before spiking mask decoding.

3.3 Spiking representation with soft-reset dynamics

We adopt leaky integrate-and-fire neurons with soft reset. Given an input current $X[t]$, the membrane potential and spike output at time step t are updated as

$$\tilde{U}[t] = \tau U[t-1] + X[t], \tag{1}$$
$$S[t] = \mathbb{I}(\tilde{U}[t] \geq V_{\text{th}}), \tag{2}$$
$$U[t] = \tilde{U}[t] - V_{\text{th}} S[t], \tag{3}$$

where τ is the membrane decay factor and V_{th} is the firing threshold. Directly simulating binary spikes through many time steps is difficult for dense prediction. Following SFA [38], we use an integer-valued spike count during training:

$$S_D^l = \text{clip}\left(\left\lfloor \frac{U^l}{V_{\text{th}}} + \frac{1}{2} \right\rfloor, 0, D \right), \qquad S_D^l \in \{0, 1, \ldots, D\}, \tag{4}$$

where U^l denotes the pre-activation membrane potential in layer l, and D is the maximum spike count. The normalization by V_{th} makes the spike count independent of the threshold scale. During back-propagation, the rounding and clipping operations use a straight-through surrogate gradient inside the clipping range and zero gradient outside it. At inference, the spike count is deterministically

expanded into a binary spike train $\{\hat{S}^l[d]\}_{d=1}^{D}$ by setting the first S_D^l bins to one:

$$\hat{S}^l[d] = \mathbb{I}(d \le S_D^l), \qquad S_D^l = \sum_{d=1}^{D} \hat{S}^l[d]. \tag{5}$$

This design keeps the model in a spike-based representation while preserving enough information for pixel-level mask prediction. In Table 3, SFA denotes this count-based training and deterministic spike-train expansion, whereas LIF denotes direct multi-step LIF simulation with the same soft-reset rule.

3.4 AVSegFormer-SNN baseline

The AVSegFormer-SNN baseline is obtained by converting AVSegFormer into an SNN and removing the MI-guided fusion objective from the full framework. It uses the same visual encoder, audio encoder, audio-query interaction, spiking transformer, and segmentation decoder as MI-SNN, but the interaction between audio and visual features is performed only by the standard AVSegFormer-style feature fusion. Formally, the baseline produces a fused representation

$$F_{\text{base}} = \Phi_{\text{SNN}}(F_v, F_a), \tag{6}$$

where Φ_{SNN} denotes the AVSegFormer-SNN fusion and decoding pipeline. This baseline is intentionally simple: it isolates the effect of spike-based representation from the effect of mutual-information alignment. As shown in the experiments, this model can learn the AVS task but suffers a large accuracy drop, indicating that spike conversion alone does not preserve the cross-modal precision required by dense segmentation.

3.5 Mutual-information-guided fusion

MI-SNN introduces an explicit dependency estimator between audio and visual features. For each frame-audio pair n and spatial location i in the visual feature map, we pair the local visual vector $f_{v,n}^{(i)}$ with the synchronized audio vector $f_{a,n}$. A discriminator T_θ scores a pair as $s_{n,i,m} = T_\theta(f_{v,n}^{(i)}, f_{a,m})$. Positive pairs use $m = n$, and negative pairs are generated by shuffling audio features within the mini-batch. The empirical Donsker–Varadhan lower bound is

$$\widehat{I}_{\text{DV}} = \frac{1}{NHW} \sum_{n,i} \left[s_{n,i,n} - \log \frac{1}{N-1} \sum_{m \neq n} \exp(s_{n,i,m}) \right]. \tag{7}$$

The auxiliary MI loss is the negative lower bound,

$$\mathcal{L}_{\text{MI}} = -\widehat{I}_{\text{DV}}, \tag{8}$$

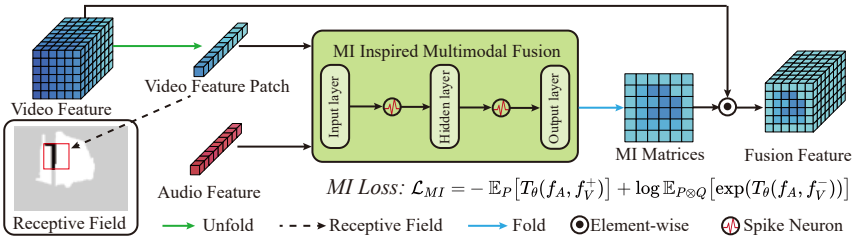

Fig. 3. MI-guided multimodal fusion. MI is estimated between global audio features and local visual features to enhance sound-related spatial regions.

so minimizing the final objective maximizes the estimated audio-visual dependency for matched samples while separating mismatched pairs. The spatial dependency map used for fusion is the positive-pair score map $A_{\mathrm{MI}} \in \mathbb{R}^{N\times 1\times H\times W}$ with entries $A_{\mathrm{MI}}[n,i] = s_{n,i,n}$. The visual representation is enhanced as

$$F_v^+ = F_v \odot (1 + \sigma(A_{\mathrm{MI}})), \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Thus, visual locations that are more dependent on the audio cue receive stronger responses before decoding.

The enhanced visual features are combined with audio-conditioned spiking queries. Let Q_a denote the spike-based audio query and K_v, V_v denote spike-based key and value projections of the visual tokens. The spiking transformer updates the query representation by

$$Q_{\mathrm{out}} = \mathrm{SN}\left(\mathrm{Attn}(Q_a, K_v, V_v)\right), \tag{10}$$

where $\mathrm{SN}(\cdot)$ denotes the spiking neuron operation. The decoder then predicts the segmentation mask from Q_{out} and F_v^+. Fig. 3 illustrates the MI-guided fusion module. Fig. 4 shows that the learned MI maps highlight regions that correspond to the sounding object, explaining why MI-SNN improves the AVSegFormer-SNN baseline.

3.6 Training objective

The segmentation output is supervised by a mask loss $\mathcal{L}_{\mathrm{seg}}$ reduced over all folded frame-audio pairs,

$$\mathcal{L}_{\mathrm{seg}} = \frac{1}{N}\sum_{n=1}^{N} \ell_{\mathrm{seg}}(\hat{Y}_n, Y_n). \tag{11}$$

The MI module provides the auxiliary loss in Eq. (8). The final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \lambda \mathcal{L}_{\mathrm{MI}}, \tag{12}$$

where λ balances segmentation supervision and MI-guided alignment. In all main experiments, the AVSegFormer-SNN baseline sets the MI term to zero, while MI-SNN uses the full objective. Algorithm 1 summarizes the training flow and explicitly shows where the MI map is inserted into the spiking AVS pipeline.

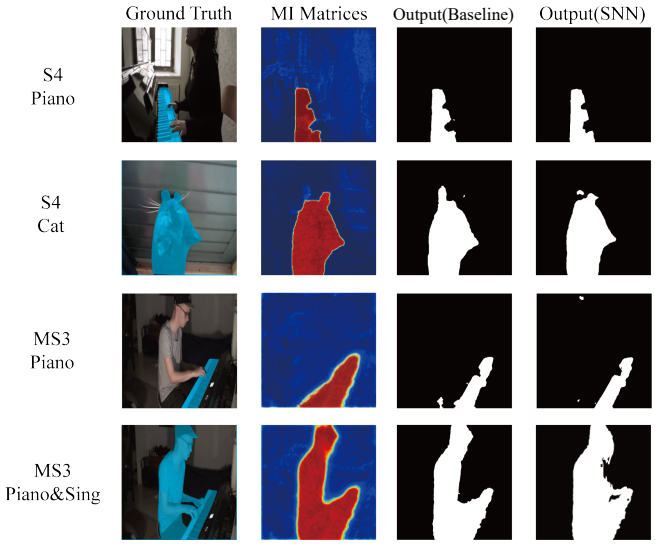

Fig. 4. Visualization of MI-guided modulation. MI maps emphasize regions that are semantically aligned with the audio signal and guide the final segmentation masks.

Algorithm 1: Training procedure of MI-SNN

Input: Video frames V, synchronized audio A, ground-truth masks Y
Output: Predicted sounding-object masks $\hat{Y}$
Extract visual features F_v with the visual encoder;
Extract audio feature F_a with the audio encoder;
Fold clip and time axes into synchronized frame-audio pairs;
Convert hidden activations into spike-based representations using SFA;
Estimate spatial MI map A_{MI} between F_a and local visual features $f_v^{(i)}$;
Enhance visual features by $F_v^+ = F_v \odot (1 + \sigma(A_{\mathrm{MI}}))$;
Fuse F_v^+ and audio-conditioned queries with the spiking transformer;
Decode the fused spike-based representation into $\hat{Y}$;
Optimize $\mathcal{L}_{\mathrm{seg}} + \lambda \mathcal{L}_{\mathrm{MI}}$;

4 Experiments

4.1 Dataset and implementation details

We evaluate on the S4 and MS3 subsets of AVSBench [40]; AVSS is outside this study's experimental scope. Following standard AVS evaluation, we report mean Intersection over Union (mIoU) and F-score. The model is trained for 50 epochs with AdamW. Frames are encoded by PVTv2, audio is converted to log-mel spectrograms and encoded by VGGish, and spiking modules use LIF-SR neurons with $V_{\mathrm{th}} = 1.0$, $\tau = 0.5$, and maximum spike count $D = 4$. Unless otherwise specified, MI-SNN uses $\lambda = 0.1$.

Table 1. Comparison on the S4 and MS3 subsets of AVSBench. The key comparison is between the AVSegFormer-SNN baseline without MI and MI-SNN with MI-guided fusion; bold marks the better result among the SNN variants.

Method	**Backbone**	**S4**		**MS3**	
		F-score	**mIoU**	**F-score**	**mIoU**
AVSBench [40]	ResNet-50	84.8	72.79	57.8	47.88
AVSegFormer [9]	PVTv2	89.9	82.06	69.3	58.36
CAVP [6]	ResNet-50	92.9	85.77	73.6	62.39
COMBO [37]	PVTv2	91.9	84.74	71.2	59.19
CQFormer [21]	PVTv2	91.2	83.62	72.7	61.05
AVSegFormer-SNN (w/o MI)	Spiking PVTv2	74.7	62.50	58.5	46.53
MI-SNN (ours)	Spiking PVTv2	**86.1**	**76.97**	**64.1**	**53.19**

4.2 Main comparison

Table 1 compares representative ANN-based AVS models with the proposed SNN variants. The key comparison is between AVSegFormer-SNN and MI-SNN. AVSegFormer-SNN obtains 74.7 F-score and 62.50 mIoU on S4, confirming that an AVSegFormer-style SNN can perform AVS but also showing a large gap to established ANN-based models. This supports our hypothesis that spike discretization weakens fine-grained audio-visual correspondence.

With mutual-information-guided fusion, MI-SNN improves S4 performance to 86.1 F-score and 76.97 mIoU, a 14.47-point mIoU gain over AVSegFormer-SNN. The task formulation is unchanged; the difference is the MI objective that guides hidden-state audio-visual alignment. On MS3, AVSegFormer-SNN obtains 58.5 F-score and 46.53 mIoU, while MI-SNN reaches 64.1 and 53.19. Thus, MI-guided fusion is also effective with multiple sound sources. Although ANN methods still benefit from continuous representations, MI-SNN gives a competitive SNN-style AVS result and shows that explicit dependency modeling is beneficial.

4.3 Qualitative analysis

Fig. 5 compares representative S4 and MS3 samples. AVSegFormer-SNN often produces incomplete masks or misses object parts, consistent with Table 1: without MI, the spike-based model lacks guidance for acoustically relevant regions. MI-SNN produces more coherent masks and better separates sounding objects from background, especially for small targets or silent distractors.

4.4 Analysis of the SNN performance gap

The gap between AVSegFormer-SNN and MI-SNN follows from the dense nature of AVS: each spatial location must be judged for acoustic relevance. Spike counts can compress differences between audio-relevant and irrelevant regions,

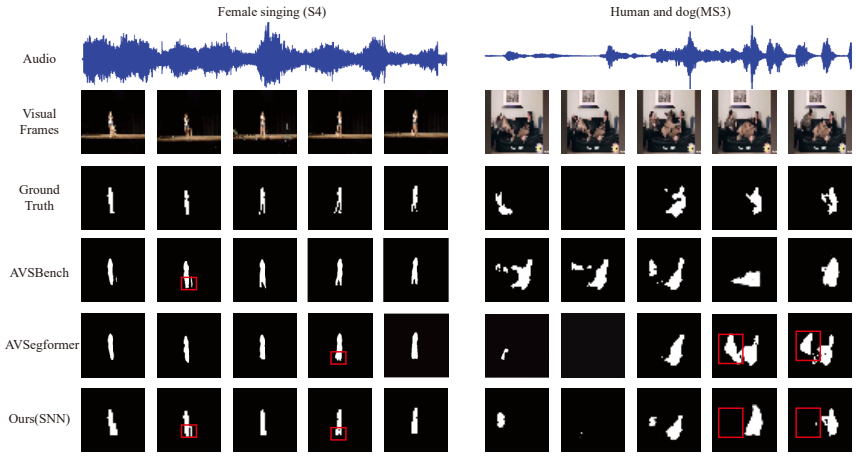

Fig. 5. Qualitative comparison on AVSBench. MI-SNN produces more complete masks than the AVSegFormer-SNN baseline and improves the localization of sounding regions.

Table 2. Ablation of MI-guided fusion on the S4 subset.

Model	MI-guided fusion	F-score	mIoU
AVSegFormer-SNN baseline	No	74.7	62.50
MI-SNN (ours)	Yes	86.1	76.97

producing coarse masks. MI-guided fusion compensates with a pre-decoding dependency map that amplifies locations with stronger audio-visual dependency, explaining the more complete masks in Fig. 5 and the MS3 gains.

4.5 Ablation of mutual information

Table 2 isolates the contribution of MI-guided fusion. Removing MI gives the AVSegFormer-SNN baseline, whose mIoU is 62.50, whereas adding MI raises it to 76.97. The gain is large because MI supplies the pixel-level sound-region dependency that spike-based features alone fail to preserve.

4.6 Effect of spike-count configuration

We further evaluate the spike-count configuration used by SFA. Table 3 shows that increasing the maximum spike count from $D = 2$ to $D = 4$ improves the F-score from 81.9 to 86.1 and mIoU from 76.56 to 76.97. Increasing D to 8 gives only a small mIoU gain, so we use $D = 4$ as the default trade-off. The direct LIF comparison clarifies that SFA is a count-based training and inference approximation rather than a separate neuron family, and the consistently higher SFA results indicate that the count-based representation is more suitable for our dense AVS decoder.

Table 3. Effect of spike-count configuration on S4.

Configuration	F-score	mIoU
SFA ($D = 2$)	81.9	76.56
SFA ($D = 4$)	86.1	76.97
SFA ($D = 8$)	86.1	77.02
LIF ($T = 2$)	80.5	76.11
LIF ($T = 4$)	80.8	76.34
LIF ($T = 8$)	81.3	76.80

4.7 Discussion

The experiments show that SNNs are feasible for multimodal AVS: AVSegFormer-SNN learns sounding-object masks, so spike-based representations can support dense audio-visual prediction. However, feasibility is not enough. Direct spike-based modeling loses cross-modal information, whereas MI-guided fusion restores an alignment signal by estimating which visual regions depend on audio. This explains the gains in Tables 1 and 2 and the clearer masks in Fig. 5.

5 Conclusion

This paper presents MI-SNN, an SNN-based AVS framework built on an AVSegFormer-SNN baseline. Directly converting the AVSegFormer-style pipeline into an SNN causes a severe drop because spike-based representations weaken fine-grained audio-visual alignment. MI-SNN addresses this with mutual-information-guided fusion: by estimating dependency between audio and spatial visual features, it enhances sound-related regions and substantially improves the SNN baseline on S4 and MS3. These results show that explicit cross-modal dependency modeling is effective for extending SNNs to multimodal dense prediction.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62373016), the Brain Science and Brain-like Intelligence Technology–National Science and Technology Major Project (Grant No. 2025ZD0217200), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB1010302), the CAS Project for Young Scientists in Basic Research (Grant No. YSBR-116), the Youth Innovation Promotion Association of CAS, the Shanghai Leading Talent Program of the Eastern Talent Plan, the Lingang Laboratory Fund (Grant Nos. LG-GG-202402-06-07 and LGL-1987-09), the Shanghai Municipal Science and Technology Project (Grant Nos. 25ZR1401370 and 25LN3200400), and the Special Support Project of Guangdong Province (Grant No. 0720240209).

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
2. Belghazi, M.I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, R.D.: MINE: Mutual information neural estimation. arXiv preprint arXiv:1801.04062 (2018)
3. Bhosale, S., Yang, H., Kanojia, D., Zhu, X.: Leveraging foundation models for unsupervised audio-visual segmentation. arXiv preprint arXiv:2309.06728 (2023)
4. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16867–16876 (June 2021)
5. Chen, T., Tan, Z., Gong, T., Chu, Q., Wu, Y., Liu, B., Yu, N., Lu, L., Ye, J.: Bootstrapping audio-visual video segmentation by strengthening audio cues. IEEE Transactions on Circuits and Systems for Video Technology (2024)
6. Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., Frazer, H., Carneiro, G.: Unraveling instance associations: A closer look for audio-visual segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26497–26507 (2024)
7. Chen, Y., Wang, C., Liu, Y., Wang, H., Carneiro, G.: CPM: Class-conditional prompting machine for audio-visual segmentation. In: European Conference on Computer Vision. pp. 438–456. Springer (2024)
8. Deng, S., Gu, S.: Optimal conversion of conventional artificial neural networks to spiking neural networks (2021), https://arxiv.org/abs/2103.00476
9. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: AVSegFormer: Audio-visual segmentation with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 12155–12163 (2024)
10. Geng, T., Wang, T., Duan, J., Cong, R., Zheng, F.: Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22942–22951 (2023)
11. Hadizadeh, H., Yeganli, S.F., Rashidi, B., Bajić, I.V.: Mutual information analysis in multimodal learning systems. In: 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 390–395. IEEE (2024)
12. Hareb, D., Martinet, J.: EvSegSNN: Neuromorphic semantic segmentation for event data. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2024). https://doi.org/10.1109/IJCNN60899.2024.10650811
13. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 131–135 (2017). https://doi.org/10.1109/ICASSP.2017.7952132
14. Hsu, W.N., Glass, J.: Multi-modal deep learning with mutual information maximization for video analysis. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 3017–3021 (2019)
15. Huang, S., Ling, R., Hui, T., Li, H., Zhou, X., Zhang, S., Liu, S., Hong, R., Wang, M.: Revisiting audio-visual segmentation with vision-centric transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8352–8361 (2025)

16. Jia, S., Zhang, T., Cheng, X., Liu, H., Xu, B.: Neuronal-plasticity and reward-propagation improved recurrent spiking neural networks. Frontiers in Neuroscience **15**, 654786 (2021). https://doi.org/10.3389/fnins.2021.654786
17. Kim, Y., Chough, J., Panda, P.: Beyond classification: directly training spiking neural networks for semantic segmentation. Neuromorphic Computing and Engineering **2**(4), 044015 (dec 2022). https://doi.org/10.1088/2634-4386/ac9b86, https://dx.doi.org/10.1088/2634-4386/ac9b86
18. Li, J., Yu, S., Wang, Y., Wang, L., Lu, H.: SeLM: Selective mechanism based audio-visual segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3926–3935 (2024)
19. Li, M., Su, N., Qu, F., Zhong, Z., Chen, Z., Tu, Z., Li, X.: VISTA: Enhancing vision-text alignment in MLLMs via cross-modal mutual information maximization. arXiv preprint arXiv:2505.10917 (2025)
20. Ling, Y., Li, Y., Gan, Z., Zhang, J., Chi, M., Wang, Y.: TransAVS: End-to-end audio-visual segmentation with transformer. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7845–7849. IEEE (2024)
21. Lv, Y., Liu, Z., Chang, X.: Consistency-queried transformer for audio-visual segmentation. IEEE Transactions on Image Processing (2025)
22. Ma, J., Sun, P., Wang, Y., Hu, D.: Stepping stones: A progressive training strategy for audio-visual semantic segmentation. In: European Conference on Computer Vision. pp. 311–327. Springer (2024)
23. Mao, Y., Zhang, J., Xiang, M., Lv, Y., Li, D., Zhong, Y., Dai, Y.: Contrastive conditional latent diffusion for audio-visual segmentation. IEEE Transactions on Image Processing **34**, 4108–4119 (2025). https://doi.org/10.1109/TIP.2025.3580269
24. Mo, S., Tian, Y.: AV-SAM: Segment anything model meets audio-visual localization and segmentation. arXiv preprint arXiv:2305.01836 (2023)
25. Nguyen, C.V.T., Nguyen, N.H.T., Le, D.T., Ha, Q.T.: Self-MI: Efficient multimodal fusion via self-supervised multi-task learning with auxiliary mutual information maximization. arXiv preprint arXiv:2311.03785 (2023)
26. Parameshwara, C.M., Li, S., Fermüller, C., Sanket, N.J., Evanusa, M.S., Aloimonos, Y.: SpikeMS: Deep spiking neural network for motion segmentation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3414–3420. IEEE (2021)
27. Poole, B., Ozair, S., Van den Oord, A., Alemi, A.A., Tucker, G.: On variational bounds of mutual information. International Conference on Machine Learning pp. 5171–5180 (2019)
28. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 292–308. Springer International Publishing, Cham (2020)
29. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4358–4366 (2018)
30. Sun, P., Zhang, H., Hu, D.: Unveiling and mitigating bias in audio visual segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7259–7268 (2024)
31. Sun, Z., Sarma, P., Sethares, W.A., Liang, Y.: Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 8992–8999 (2020)

32. Wang, Q., Zhang, T., Han, M., Wang, Y., Zhang, D., Xu, B.: Complex dynamic neurons improved spiking transformer network for efficient automatic speech recognition. Proceedings of the AAAI Conference on Artificial Intelligence **37**(1), 102–109 (2023). https://doi.org/10.1609/aaai.v37i1.25107
33. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 568–578 (October 2021)
34. Wang, Y., Sun, P., Li, Y., Zhang, H., Hu, D.: Can textual semantics mitigate sounding object segmentation preference? In: European Conference on Computer Vision. pp. 340–356. Springer (2024)
35. Wang, Z., Yang, Q., Shi, L., Yu, J., Liang, Q., Li, F., Xiang, S.: AVESFormer: Efficient transformer design for real-time audio-visual segmentation. arXiv preprint arXiv:2408.01708 (2024)
36. Xuan, H., Wu, Z., Yang, J., Jiang, B., Luo, L., Alameda-Pineda, X., Yan, Y.: Robust audio-visual contrastive learning for proposal-based self-supervised sound source localization in videos. IEEE transactions on pattern analysis and machine intelligence **46**(7), 4896–4907 (2024)
37. Yang, Q., Nie, X., Li, T., Gao, P., Guo, Y., Zhen, C., Yan, P., Xiang, S.: Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27134–27143 (2024)
38. Yao, M., Qiu, X., Hu, T., Hu, J., Chou, Y., Tian, K., Liao, J., Leng, L., Xu, B., Li, G.: Scaling spike-driven transformer with efficient spike firing approximation training. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2973–2990 (2025). https://doi.org/10.1109/TPAMI.2025.3530246
39. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2236–2246 (2018)
40. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., et al.: Audio-visual segmentation with semantics. International Journal of Computer Vision **133**(4), 1644–1664 (2025)