

# Supplementary Materials for CACE-Net: Co-guidance Attention and Contrastive Enhancement for Effective Audio-Visual Event Localization

Xiang He\*  
Xiangxi Liu\*  
Yang Li\*

Brain-inspired Cognitive Intelligence Lab,  
Institute of Automation, Chinese Academy of Sciences  
Beijing, China  
{hexiang2021,liuxiangxi2024,liyang2019}@ia.ac.cn

Guobin Shen

Brain-inspired Cognitive Intelligence Lab,  
Institute of Automation, Chinese Academy of Sciences  
Beijing, China  
Center for Long-term Artificial Intelligence  
Beijing, China  
shenguobin2021@ia.ac.cn

Xin Yang<sup>†</sup>

Institute of Automation, Chinese Academy of Sciences  
Beijing, China  
xin.yang@ia.ac.cn

Dongcheng Zhao

Brain-inspired Cognitive Intelligence Lab,  
Institute of Automation, Chinese Academy of Sciences  
Beijing, China  
Center for Long-term Artificial Intelligence  
Beijing, China  
zhaodongcheng2016@ia.ac.cn

Qingqun Kong<sup>†</sup>

Brain-inspired Cognitive Intelligence Lab,  
Institute of Automation, Chinese Academy of Sciences  
Beijing, China  
qingqun.kong@ia.ac.cn

Yi Zeng<sup>†</sup>

Brain-inspired Cognitive Intelligence Lab,  
Institute of Automation, Chinese Academy of Sciences  
Beijing, China  
Center for Long-term Artificial Intelligence  
Beijing, China  
Key Laboratory of Brain Cognition and Brain-inspired  
Intelligence Technology, CAS  
Shanghai, China  
yi.zeng@ia.ac.cn

## ACM Reference Format:

Xiang He, Xiangxi Liu, Yang Li, Dongcheng Zhao, Guobin Shen, Qingqun Kong, Xin Yang, and Yi Zeng. 2024. Supplementary Materials for CACE-Net: Co-guidance Attention and Contrastive Enhancement for Effective Audio-Visual Event Localization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28-November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3664647.3681503>

\*Authors contributed equally.

<sup>†</sup>Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3681503>

## 1 Analysis and Discussion

In this section, we explore several factors that influence the experimental results, focusing on the temporal encoder, the hyperparameter settings, and we conclude with a summary of the limitations of the work and directions for future research.

**The effect of temporal encoder.** We have shown the framework of our proposed Audio-Visual Co-guidance Network in Figure 2 in the main paper. In particular, it is noted that the visual and auditory modal features, after being processed by the audio visual co-guidance attention (AVCA), also need to be processed by the temporal encoder before they can be fed into the Interactive Modal Correlation Module. This is because it is crucial to consider temporal features when solving the task of audio visual event localization, as audio visual content is not static but dynamically unfolds over time. This dynamism means that the information and relationships in the audio visual data are continuously changing and depend on the content before and after in time.

In order to determine the most appropriate type of temporal encoder for the audio visual event localization task, we compare the effects of three different temporal encoders as well as the effects of not using the temporal encoder. Specifically, these three temporal

**Table S1: The effect of different temporal encoders.**

Method	Accuracy(%)
w/o temporal encoder	76.39
w/ unidirectional LSTM	77.04
w/ SNN	77.11
w/ bidirectional LSTM	<b>80.80</b>

encoders include a unidirectional long short-term memory network (LSTM), a spiking neural network (SNN), and a bidirectional LSTM. For the detailed implementation of SNN, we use a network architecture that includes a single hidden layer with a time step of 15. The neurons employed are Leaky Integrate-and-Fire (LIF) neurons [1] and the experimental results are shown in Table S1.

It is evident that compared to not using a temporal encoder, where the model's accuracy is 76.39%, employing a temporal encoder to exploit the temporal features within the modality can improve the performance of the network. For the unidirectional temporal networks, where inputs are propagated forward in time, the spiking neural network marginally outperforms the LSTM and improves the model performance by 0.7% compared to not using the temporal encoder. This demonstrates the effectiveness of SNN's ability to exploit temporal feature information in the task of audio visual event localization by modelling the mechanism of spike transmission between neurons. Following this, when the temporal encoder uses a bidirectional LSTM, the model obtains an optimal performance of 80.80%, which indicates before-and-after temporal information facilitates the judgment of the results in the intermediate moments in the task of audio visual event localization. Specifically, the bidirectional LSTM is able to capture the temporal dependencies in the video in a more comprehensive way by considering both past and future contextual information, thus effectively integrating the information from the previous and subsequent frames. This integration of before-and-after information is particularly critical for accurate recognition and localization of audio visual events, since the context of an event is usually not limited to a single instant but involves a continuous dynamic process.

**Hyperparameters setting.** In our proposed framework for audio-visual co-guidance networks, there are three key hyperparameters involved: the parameter  $\beta$  used to regulate the guiding effect of audio on visual, the parameter  $\psi$  used to modulate the proportion of visual-guidance on audio, and the fusion coefficient  $\lambda$  used in background-event contrast enhancement. For  $\beta$ , we set it to 0.4 by default; for the effect of  $\lambda$ , we show the effect of different parameter values on the results in detail in Table 2 in the main paper. Meanwhile, for  $\psi$  in Audio-Visual Co-guidance Attention (AVCA), its specific impact on network performance is also illustrated in Table S2. Obviously, compared with  $\psi = 0$ , i.e., audio-only guidance of visual in co-guidance attention, an appropriate selection of  $\psi$  (e.g., set to 0.3 or 0.45) can effectively improve network performance. However, if  $\psi$  is set too small, the desired positive effect may not be achieved due to insufficient strength.

**Limitation and future work.** Our research focuses on the task of audiovisual event localization and analyzes the difficulties and challenges of the task. Although we propose effective solutions,

**Table S2: Ablation experiments for audio-visual co-guidance attention. The experiment demonstrates the effect of the parameter  $\psi$  on the results in the Visual-guided enhancement audio feature.**

Methods	Accuracy
AVCA-Visual-only $\psi = 0$	78.83
w/ AVCA $\psi = 0.15$	78.71
w/ AVCA $\psi = 0.3$	<b>80.30</b>
w/ AVCA $\psi = 0.45$	79.93

**Table S3: Model performance on the UnAV-100 datasets.**

AVCA	BECE	0.5	0.6	0.7	0.8	0.9	Avg.
-	-	49.3	45.0	39.5	<b>32.9</b>	<b>21.6</b>	46.8
✓	-	49.8	45.1	40.0	32.6	21.3	47.1
✓	✓	<b>50.1</b>	<b>45.4</b>	<b>40.2</b>	32.7	21.2	<b>47.5</b>

the AVE dataset itself has some limitations. We note that some videos in the dataset have problems with labeling, e.g., events in the videos may appear intermittently while the labels are labeled as continuous occurrences, which may lead to model predictions being misclassified as errors due to mislabeling even if they are consistent with human observations.

In addition, there is only one instance in each video in the AVE dataset. This setup is inconsistent with the reality of natural videos, which often contain multiple audio visual events of different categories. Accomplishing this task on unconstrained video datasets that contain more dense events will be more relevant to real-world application scenarios. Therefore, applying our method to larger and more event-dense datasets, such as the Untrimmed Audio-Visual (UnAV-100) dataset [2], is a direction worth further exploration.

## 2 Generalization on more datasets

For audio-visual event localization tasks, previous related studies were conducted exclusively on the AVE dataset. However, further validation of our method in real-world scenarios is necessary. Therefore, we select the UnAV-100 dataset [3], a large-scale untrimmed audio-visual dataset, to evaluate the effectiveness and generalizability of our proposed methods. UnAV-100 dataset contains multiple categories of audio-visual events, often occurring simultaneously within a video, just as they do in real-world scenarios.

Specifically, we employ the efficient encoder provided by Geng et al. [3]. to validate our proposed Audio-Visual Co-Guidance Attention (AVCA) and Background-Event Contrast Enhancement (BECE) methods. We conduct experiments on the UnAV-100 dataset, with results shown in Table S3. We report the mean Average Precision (mAP) at the tIoU thresholds [0.5:0.1:0.9] and the average mAP at thresholds [0.1:0.1:0.9]. It can be observed that the average mAP with AVCA reaches 47.1%, which is better compared to the baseline. Further integration of BECE enables the model to achieve optimal performance, demonstrating the effectiveness of our methods and their generalizability on the new dataset.

**Table S4: Results on the AVE datasets. Bolding represents the best results, underlining represents results in our paper.**

Augmentation Method	Accuracy	BECE loss weights	Accuracy
Channel Random Mask	79.20%	$\lambda_1 = 0.0$	80.30%
Feature Mixup	78.36%	$\lambda_1 = 0.5$	80.42%
Random Noise $\sigma = 1.0$	79.50%	$\lambda_1 = 1.0$	<u>80.80%</u>
Random Noise $\sigma = 0.1$	<b>80.80%</b>	$\lambda_1 = 3.0$	<b>80.87%</b>
Random Noise $\sigma = 0.05$	80.72%	$\lambda_1 = 5.0$	80.30%

(a) Method comparison      (b) Coefficient comparison

**Table S5: Ablation experiments of fine-tuning.**

Feature Extractor	Method	Supervised
VGG19	w/o Fine-tuning	79.65
	w/ Fine-tuning	80.05
ResNet50	w/o Fine-tuning	81.56
	w/ Fine-tuning	82.36

### 3 More ablation experiment analysis

We conduct additional experimental analyses and included three more experiments: 1) The impact of different data augmentation methods on event-background prediction. We explore three data augmentation techniques: channel random masking (zeroing), feature mixup, and random Gaussian noise [ $X \sim \mathcal{N}(\mu = 0, \sigma^2)$ ], with results shown in Table S4. Gaussian noise augmentation proved to be the most effective, which we believe is due to the similar characteristics of event and background features, and Gaussian noise helps to highlight the subtle differences between them. 2) Targeted fine-tuning on the existing encoder. We use VGG19 as the visual encoder to re-extract features, and the experimental results are shown in Table S5. The results indicate that targeted fine-tuning is effective across different encoders. 3) The effect of loss function weights on the results. In equation 12, we set the first two terms as the base loss function, with a fixed weight of 1. For contrast enhancement loss, we set the coefficient  $\lambda_1$ , with results presented in Table S4. It shows that a larger weight for contrastive loss may yield better results, but weights that are too large or too small are not suitable.

### 4 Complexity of our method

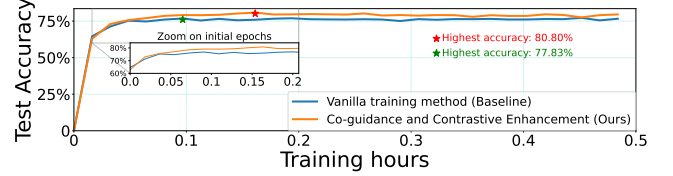
Specifically, our approach requires the incorporation of an audio-visual co-guidance attention module and a contrast enhancement projection layer, which adds additional computational complexity. According to our experiments, this extra computation is affordable and does not significantly impact inference speed while improving network performance, as shown in Table S6. Additionally, as illustrated in Figure S1, our method achieves better results than existing approach under the same training duration.

### 5 Reproducibility

Our experiments were implemented based on Pytorch [4], and for the SNN component, we chose the open source framework

**Table S6: Comparison of complexity metrics**

Method	Params	Memory	FLOPs	Inference Time	Accuracy
Vanilla	12.58M	3.88G	1.27G	1.35s	77.83%
Ours	13.12M	4.29G	1.66G	1.37s	80.80%

**Figure S1: Variations of accuracy with the training time.**

BrainCog [5] to implement the SNN with for conducting the experiments. All source codes and training scripts are provided in the Supplementary Material.

### References

- [1] Peter Dayan and Laurence F Abbott. 2005. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press.
- [2] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 2023. 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 (CVPR)*. 22942–22951.
- [3] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 2023. 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*. 22942–22951.
- [4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
- [5] Yi Zeng, Dongcheng Zhao, Feifei Zhao, Guobin Shen, Yiting Dong, Enmeng Lu, Qian Zhang, Yinqian Sun, Qian Liang, Yuxuan Zhao, et al. 2023. BrainCog: A spiking neural network based, brain-inspired cognitive intelligence engine for brain-inspired ai and brain simulation. *Patterns* 4, 8 (2023).