Master's Thesis Presentation for the 145th Dissertation Evaluation

A Study on CLIP-Based Global Context Regulation for Audio-Visual Event Localization

음향-영상 이벤트 위치 추정을 위한 CLIP 기반의 컨텍스트 교정 방법에 관한 연구

Sang-Rak Lee

tkdfkr1503@gmail.com

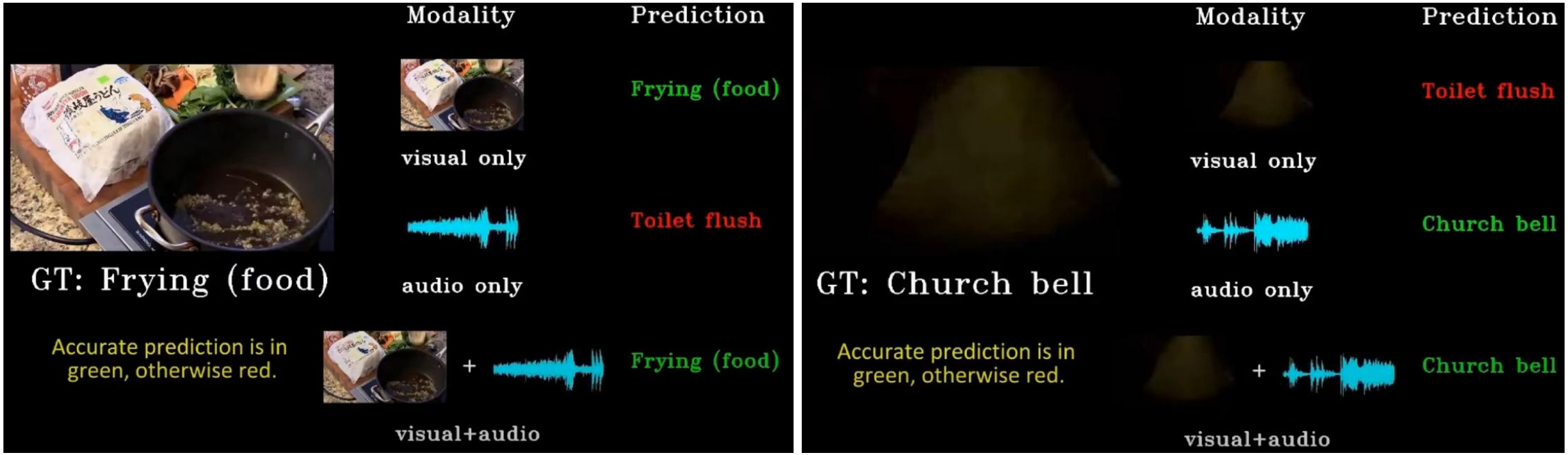
27th May, 2025

Contents

Introduction

Audio-Visual Event Localization

- AI technologies for video analysis that combine audio and visual modalities have been rapidly advancing, with multi-modality approaches demonstrating superior performance compared to single-modality methods [1].
- Audio-visual event localization (AVEL) has emerged as a significant research topic in AI-driven video analysis with advances in multi-modal technology and is applied in video surveillance, multimedia content analysis, and healthcare.



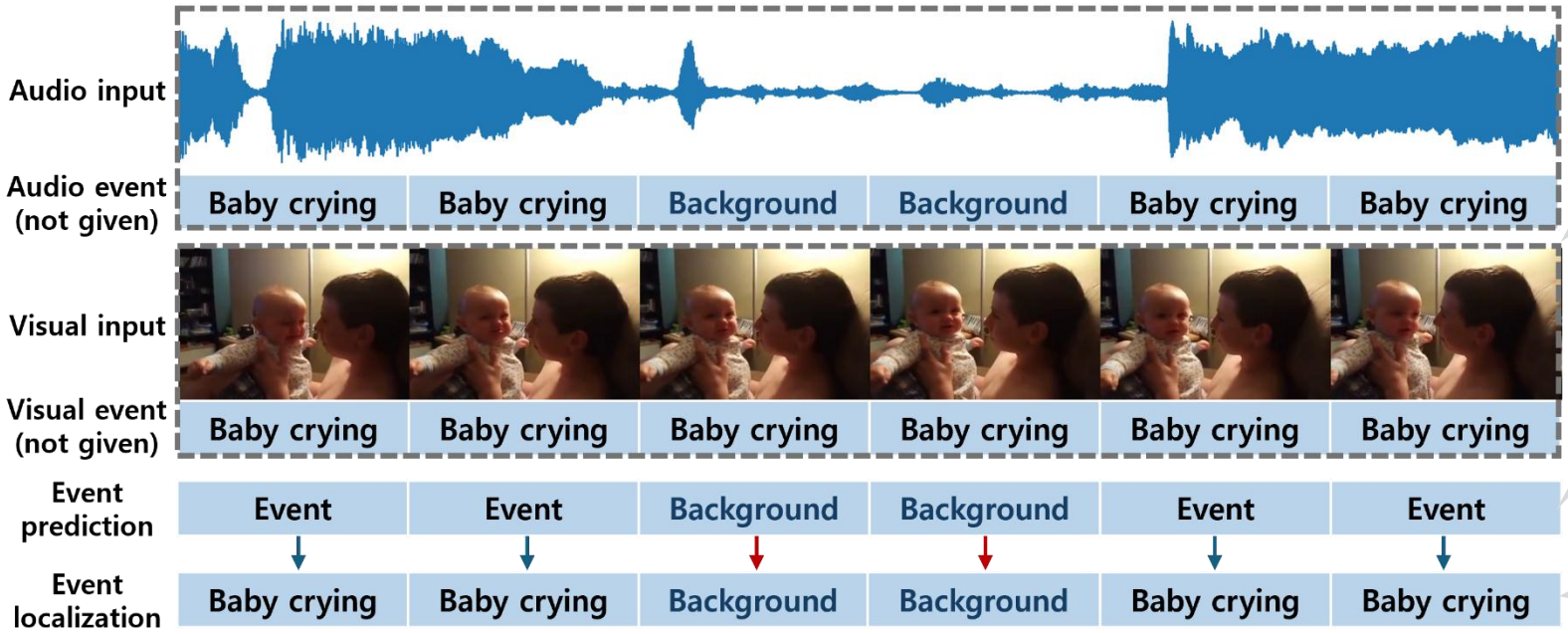
Examples of audio-visual events

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Dec. 5-10, 2016. SoundNet: Learning sound representations from unlabeled video. In Proceedings of Ad- vances in Neural Information Processing Systems, Vol. 29. Curran Associates, Inc., Barcelona, Spain, 892–900.

Introduction

Objective and Standard Procedure

- AVEL refers to leveraging audio and visual modalities to identify and precisely localize audio-visual events that occur simultaneously across both modalities over a temporal duration [2].
- In AVEL, a video is divided into non-overlapping clips, and features are extracted from visual frames and corresponding audio spectrograms as inputs to predict each segment as either an event category or background.



The audio in the third and fourth segments is labeled as ``Background (non-event),'' while the visual component is labeled as ``Baby crying.‘’

In the first step, the model predicts whether an event has occurred (Event prediction)

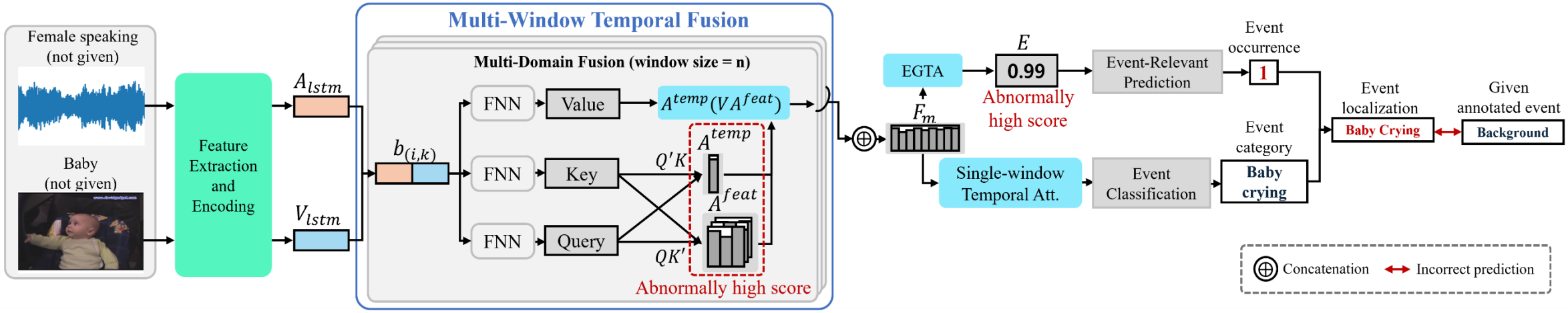
In the second step, it applies the predicted event category (Given annotated event).

AVEL example using proposed method.

Introduction

Problem and Strategy

- One challenge in AVEL is that when audio and visual contexts are inconsistent and this information is clearly present, the significance values of each modality are measured as high, causing the model to predict an event instead of the background.
- In this thesis, A post-processing method was applied to adjust the scores using time-wise cosine similarity from the initial audio and visual features, preserving the input data while preventing abnormally high event-relevant scores.



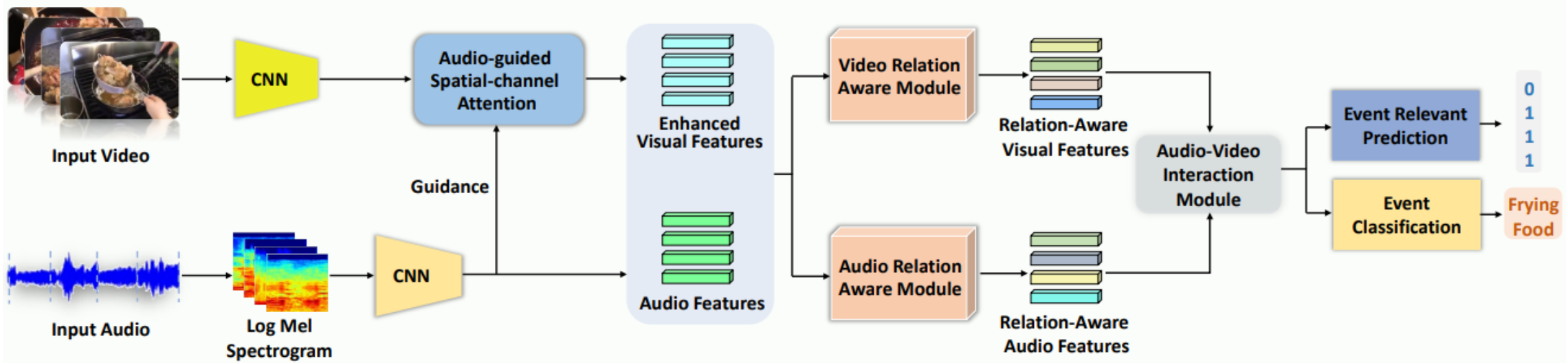
The audio modality receives “Female speaking,” while the visual modality receives “Baby.” While the model should predict background when inputs belong to different class categories, insufficient data to learn inconsistent patterns leads to a high attention score, causing the event-relevant score to incorrectly suggest an event.

Illustration of the process by which clearly inconsistent information affects predictions.

Related Work

Cross-Modal relation-aware networks for audio-visual event localization [3]

- A Cross-Modal Relation-Aware Network (CMRAN) is proposed to enhance audio-visual event localization by utilizing an audio-guided spatial-channel attention module and a relation-aware module, effectively capturing both intra- and inter-modality relations.
- Despite utilizing a relation-aware module that enhances the intra- and inter-modality information of both audio and visual, CMRAN tends to increase the significance value when the audio and visual information is inconsistent and clearly presented.



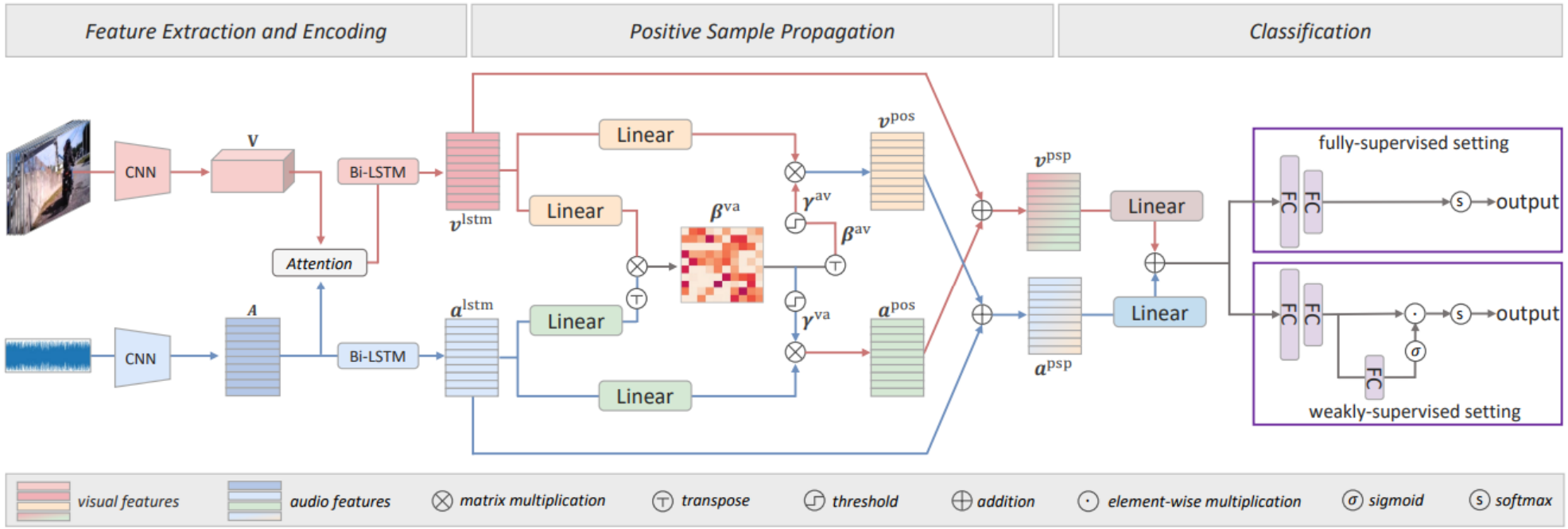
The architecture of CMRAN

[3] Xu, Haoming, et al. "Cross-modal relation-aware networks for audio-visual event localization." Proceedings of the 28th ACM International Conference on Multimedia. 2020..

Related Work

Positive sample propagation along the audio-visual event line [4]

- A Positive Sample Propagation (PSP) module is proposed to enhance audio-visual event localization by selectively aggregating highly relevant audio-visual pairs and filtering out irrelevant connections, improving classification performance.
- The PSP module performed joint representation learning for audio and visual modalities, but it still lacked a method to prevent the significance value from increasing in inconsistent cases, leading to erroneous prediction.



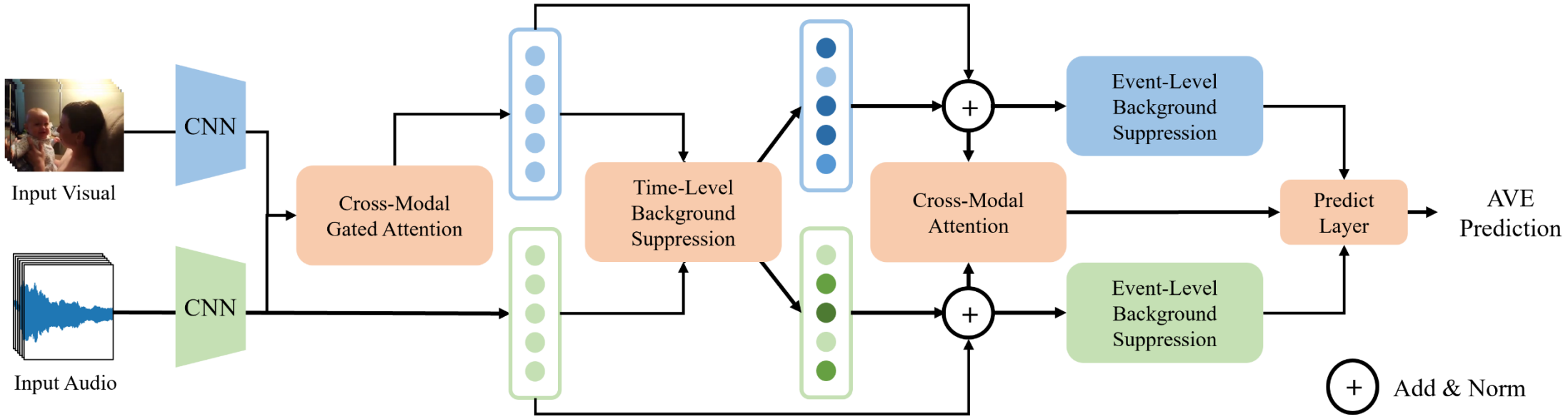
The architecture of PSP

[4] Zhou, Jinxing, et al. "Positive sample propagation along the audio-visual event line." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Related Work

Cross-Modal background suppression for audio-visual event localization [5]

- The Cross-Modal Background Suppression (CMBS) model proposes a cross-modal background suppression network to enhance event-relevant focus and suppress noise using time-level, event-level schemes, and gated attention.
- Despite effectively suppressing background noise, CMBS tends to make erroneous predictions by predicting the background as an event in situations where inconsistent information from audio and visual modalities containing different event is presented.



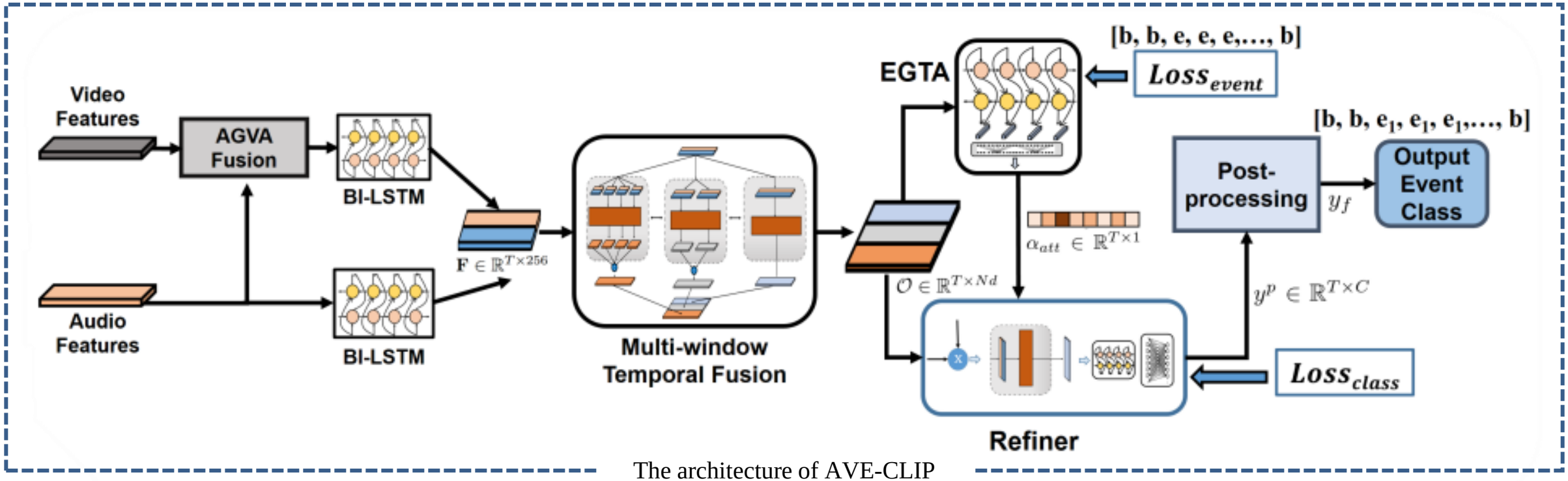
The architecture of CMBS

[5] Xia, Yan, and Zhou Zhao. "Cross-modal background suppression for audio-visual event localization." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022..

Related Work

AVE-CLIP: AudioCLIP-Based multi-window temporal transformer for audio visual event localization [6]

- AVE-CLIP, a framework that integrates AudioCLIP [7] encoders with a multi-window temporal fusion, enhances audio-visual event localization through multi-scale feature fusion and temporal refinement.
- Although AVE-CLIP enhances representations through contrastive learning of audio and visual data, it has the drawback that significance values are still measured high when inconsistent cases are provided as input.



The architecture of AVE-CLIP

[6] Mahmud, Tanvir, and Diana Marculescu. "Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Related Work

Common Drawback and Solution

- Conventional methods have the issue of measuring the significance values of each modality as high when audio-visual contexts are inconsistent and this information is clearly present, leading to erroneous predictions by predicting “Event” instead of “Background.”
- To address this issue, a CLIP-based global context regulation method is proposed, utilizing a pre-trained AudioCLIP [7] encoder to introduce an effective post-processing approach that performs well even in situations with limited inconsistent cases for training.

Method	Venue	Summary	Drawback
AVE-CLIP [6]	WACV, 2023	Conventional methods used pre-trained models for each modality separately, making multimodal processing inefficient. To overcome this, the study proposes AVE-CLIP, integrating AudioCLIP with a Multi-Window Temporal Transformer for diverse temporal scales.	Although AVE-CLIP enhances representations through contrastive learning of audio and visual data, it has the drawback that significance values are still measured high when inconsistent cases are provided as input.
CMBS [5]	CVPR, 2022	Real-world videos often have inconsistencies and background noise. The study proposes a noise removal method using one modality to reduce noise in the other at both time and event levels.	Despite effectively suppressing background noise, CMBS tends to make erroneous predictions by predicting the background as an event in situations where inconsistent information from audio and visual modalities containing different event is presented.
PSP [4]	CVPR, 2021	To enhance video content understanding, accurately identifying audio-visual pairs is key. The study proposes Positive Sample Propagation (PSP), which measures similarity among pairs and leverages closely related ones for better performance.	The PSP module performed joint representation learning for audio and visual modalities, but it still lacked a method to prevent the significance value from increasing in inconsistent cases, leading to a erroneous prediction that misclassified the background as an event.
CMRAN [3]	ACM-MM, 2020	Audio can guide attention to relevant visuals, reducing background noise. The study proposes an Audio-guided Spatial-Channel Attention module and a Relation-Aware module to connect images and audio.	Despite utilizing a relation-aware module that enhances the intra- and inter-modality information of both audio and visual, CMRAN tends to increase the significance value when the audio and visual information is inconsistent and clearly presented.

[3] Xu, Haoming, et al. "Cross-modal relation-aware networks for audio-visual event localization." Proceedings of the 28th ACM International Conference on Multimedia. 2020.
[4] Zhou, Jinxing, et al. "Positive sample propagation along the audio-visual event line." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[5] Xia, Yan, and Zhou Zhao. "Cross-modal background suppression for audio-visual event localization." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[6] Mahmud, Tanvir, and Diana Marculescu. "Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Proposed Method

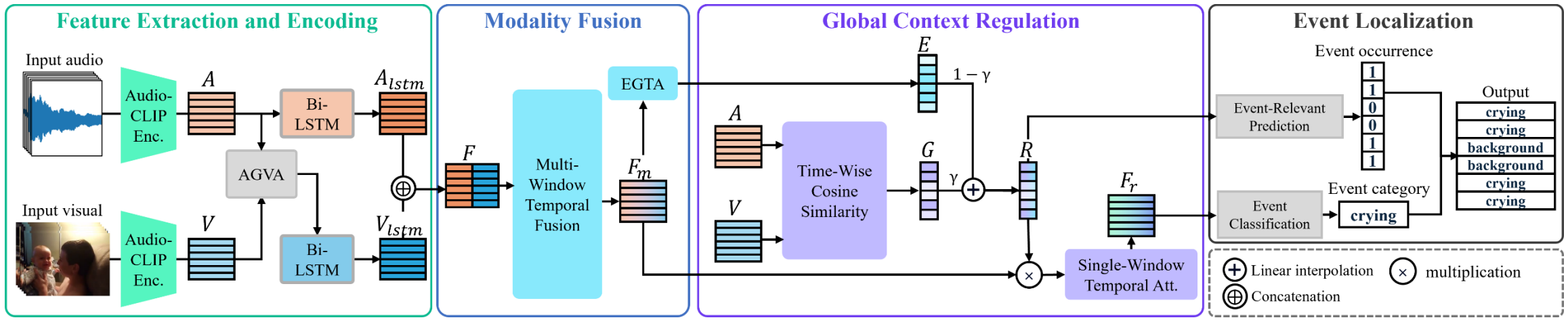
Notation and Definition

- ✓ A set of time segments : t $\in \{1, \dots, T\}$
- ✓ The corresponding audio and visual features : $A_t \in R^{d_a}$ and $V_t \in R^{d_v}$
- ✓ A fused feature generated by concatenating audio and visual features extracted using Bi-LSTM : F
- ✓ The fused feature generated through the multi-window temporal fusion (MWTF) module using F as input : F_m
- ✓ The attention vector generated by the event-guided temporal attention (EGTA) module : E
- ✓ Time-wise cosine similarity feature vectors generated by the GCR module using A and V as inputs : G
- ✓ The result of linear interpolation between E and G, a refined attention vector : R
- ✓ The weighting vector for linear interpolation : γ
- ✓ The number of event categories (including background) : C
- ✓ The prediction at time segments t : $y_t = \{y_t^k | y_t^k \in \{0,1\}, k = 1, \dots, C, \sum_{k=1}^{C} y_t^k = 1\}$
- ✓ Fully-supervised Setting : $Y^f \in R^{T \times C}, \hat{Y} = \{y_1, y_2, \dots, y_T\}$

Proposed Method

Procedural Overview

- The proposed method introduces a CLIP-based global context regulation mechanism that calculates time-wise cosine similarity between audio and visual features, effectively refining event-relevant scores for more accurate audio-visual event predictions.
- Given audio A and visual V as inputs, each is processed separately through a Bi-LSTM before being concatenated to form the fused feature F, which then undergoes modality fusion and global context regulation to generate the model prediction $\hat{Y}$.



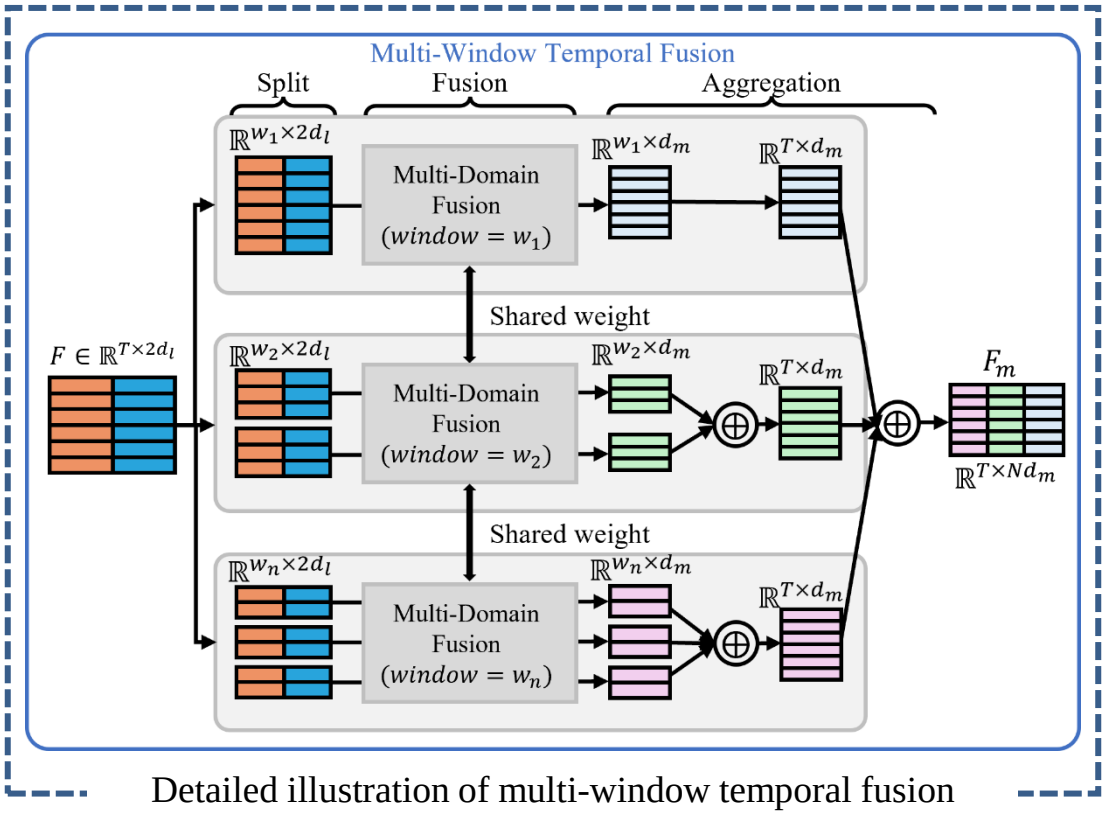
The overall pipeline of the proposed method consists of four main parts: feature extraction and encoding, modality fusion, global context regulation, and event localization.

Overview of proposed method

Proposed Method

Multi-Window Temporal Fusion

- The multi-window temporal fusion (MWTF) module utilizes multi-domain attention across various temporal scales to effectively aggregate global and local temporal information, enabling enhanced feature fusion.
- The MWTF module takes the fused feature F as input, splits it at different window scales, applies attention at the feature and time levels to perform modality fusion, and generates a fused feature F_m.

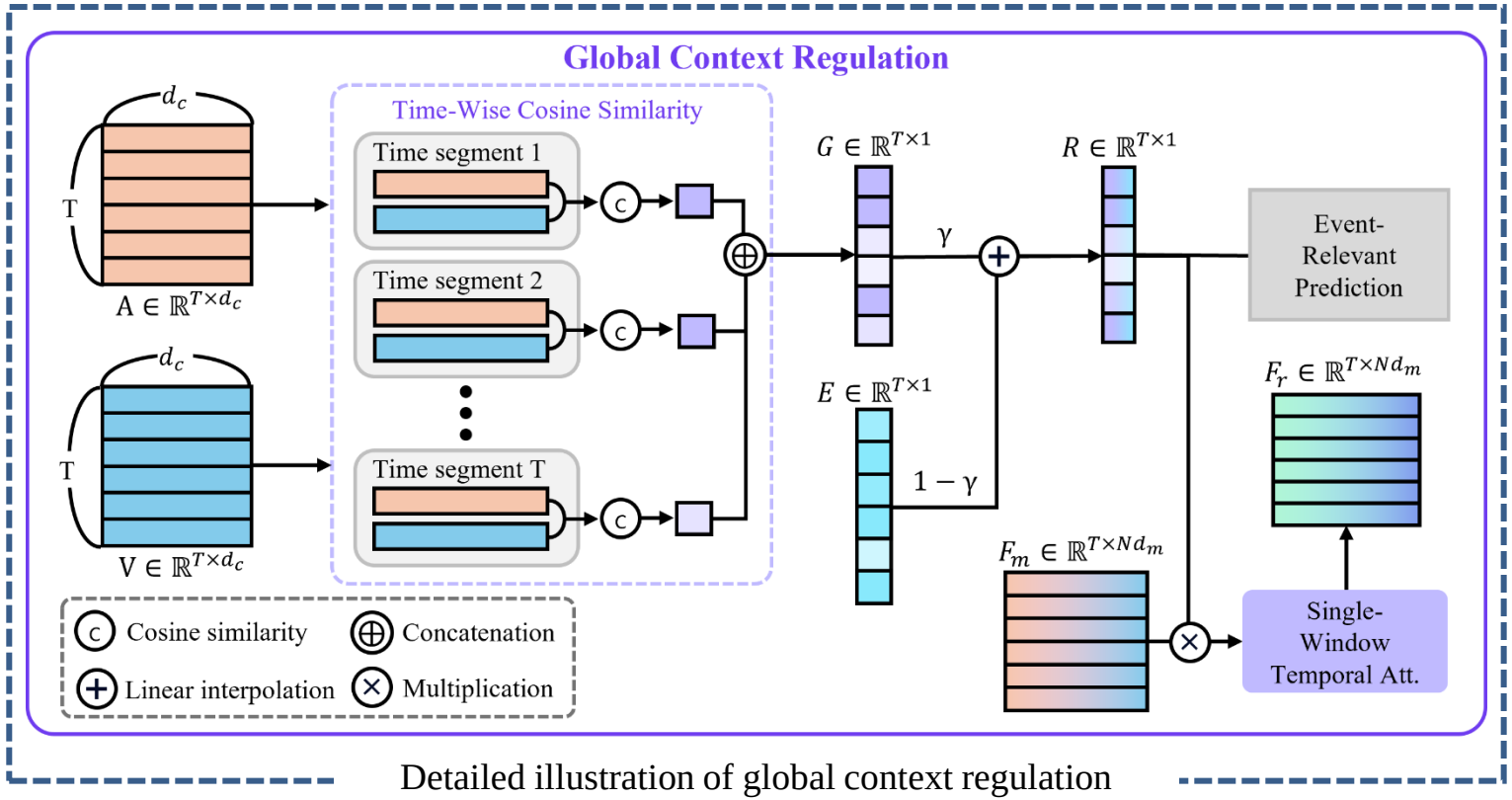


Detailed illustration of multi-window temporal fusion

Proposed Method

Global Context Regulation

- The global context regulation (GCR) is proposed to mitigate erroneous predictions by addressing the increase in significance values of each modality and the resulting rise in event-relevant scores when clear but inconsistent information is present as input.
- The GCR module takes the initial feature maps A and V as inputs, calculates the time-wise cosine similarity to generate G, and regulates the event-relevant score E from the EGTA module to generate the advanced attention vector R.

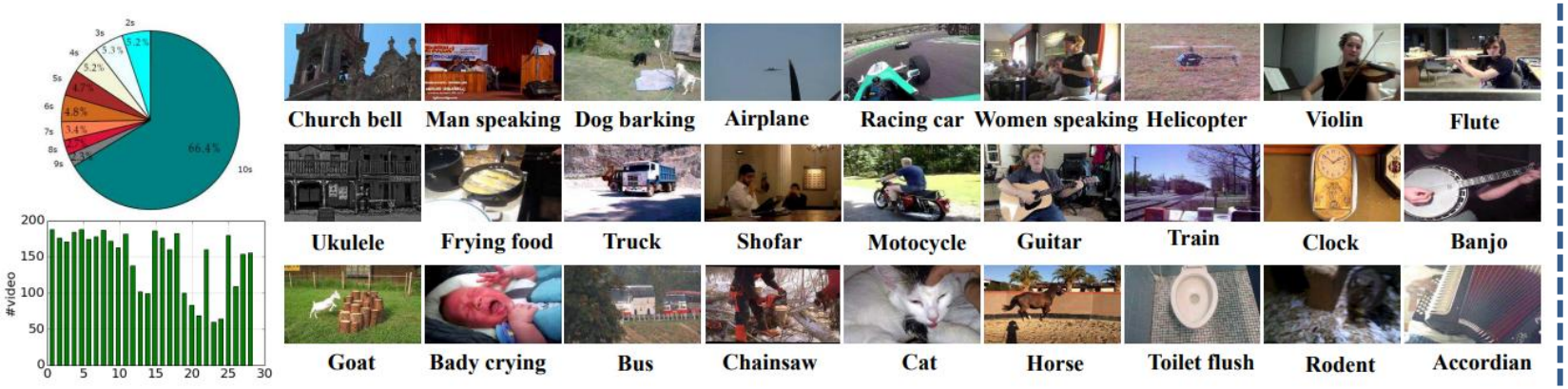


Detailed illustration of global context regulation

Experimental Settings

Dataset

- AVE [2] dataset consists of 4,143 videos, each 10 seconds long, labeled at a per-second level.
- There are a total of 28 event categories encompassing various life(in-the-wild) events.
- Each video contains at least one 2-second event, with 60 to 188 videos per event category.



The AVE dataset

Statistic	Value
Total videos	4,143
Modalities	Audio, Visual
Train/Validation/Test split	2,485 / 829 / 829 (10 random splits)
Number of classes	28
Video length	10 s
Class distribution (min/max)	60 / 188

Statistical information of the AVE dataset

Cross Validation

- Number of Iterations: 10
- Data Split Ratio (Train: Validation: Test) = 0.6 : 0.2 : 0.2

Experimental Settings

Hyperparameter Settings

- The experimental setup was based on CMRAN, which has been widely used as a baseline for other models.

Model	Epoch	Learning rate	Weight decay	Optimizer	Event classification loss	Event-Relevant prediction loss
Proposed	200	5e-4	0.5 at epochs 10, 20 and 30.	Adam	cross-entropy	binary cross-entropy
Ave-clip [6]	200	-	-	-	cross-entropy	binary cross-entropy
CMBS [5]	200	7e-4	0.5 at epochs 10, 20 and 30.	Adam	cross-entropy	binary cross-entropy
PSP [4]	200	1e-3	-	Adam	cross-entropy	binary cross-entropy
CMRAN [3]	200	5e-4	0.5 at epochs 10, 20 and 30.	Adam	cross-entropy	binary cross-entropy

Evaluation Measures

- Models for AVEL were evaluated using the classification accuracy as the evaluation measure

$$Accuracy = \frac{TP + TN}{TP + TN + FP + RN}$$

Experimental Results

Performance Comparison

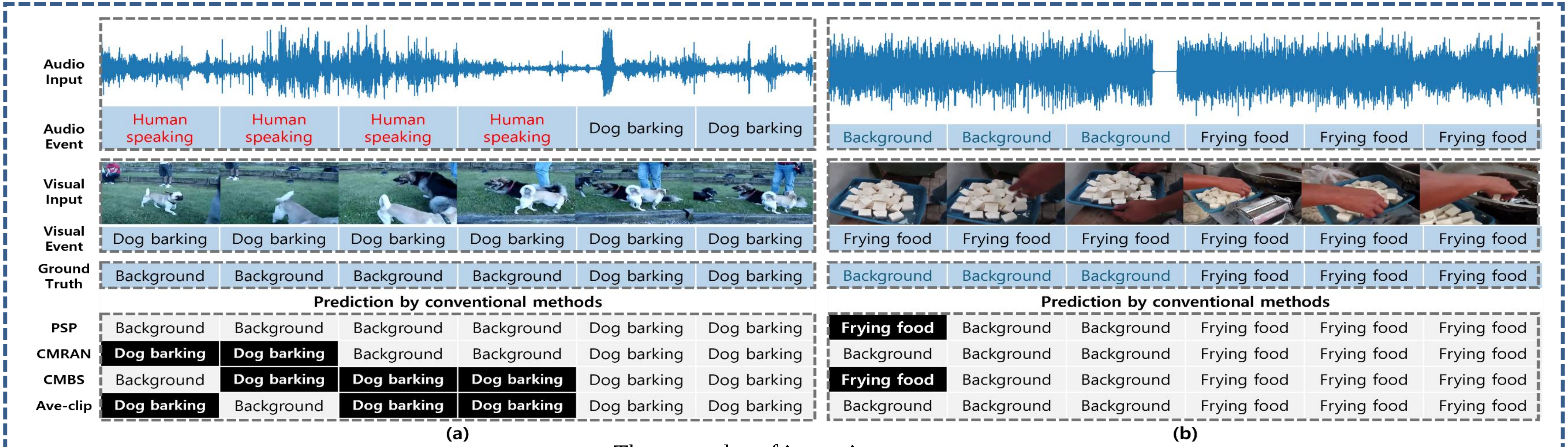
- The proposed method achieved state-of-the-art performance in event localization and demonstrated balanced improvements in both event-relevant prediction and event classification.

Model	Event Localization	Event-Relevant Prediction	Event Classification	Rank
	Accuracy (%)			
Proposed	**77.70±1.28**	**63.06±1.66**	**88.69±1.55**	**1**
Ave-clip [6]	76.77±1.32	61.70±1.19	86.99±0.90	2
CMBS [5]	76.63±1.39	59.46±1.56	86.95±1.52	3
PSP [4]	74.00±1.23	59.31±2.07	82.04±2.11	5
CMRAN [3]	75.58±1.03	57.76±2.54	87.91±0.61	4

Experimental Results

In-depth Analysis

- The key difference is that in (a), the first four audio segments contain learned information “Human speaking,” whereas in (b), the first three audio segments consist of unlearned factory noise.
- It was observed that when learned information from different classes is contained in both modalities as input and this information is clearly presented, the model tends to make erroneous predictions.

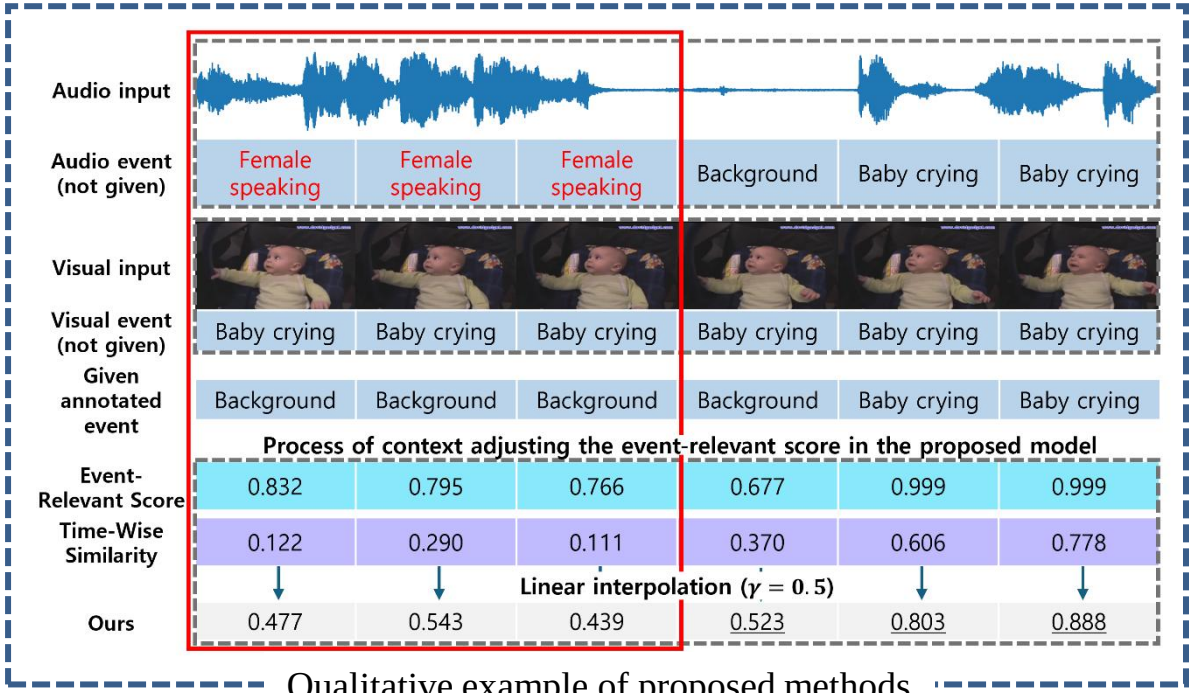


The examples of inconsistent cases

Discussion

Qualitative Analysis and Failure Analysis

- In qualitative analysis, the proposed GCR method demonstrates its ability to adjust abnormally high event-relevant scores when inconsistent information is clearly present, enabling accurate background prediction.
- A failure analysis was conducted, where "Minimum γ" denotes the minimum coefficient required for the linear interpolation of the event-relevant score and time-wise similarity to regulate γ to Background.



Qualitative example of proposed methods

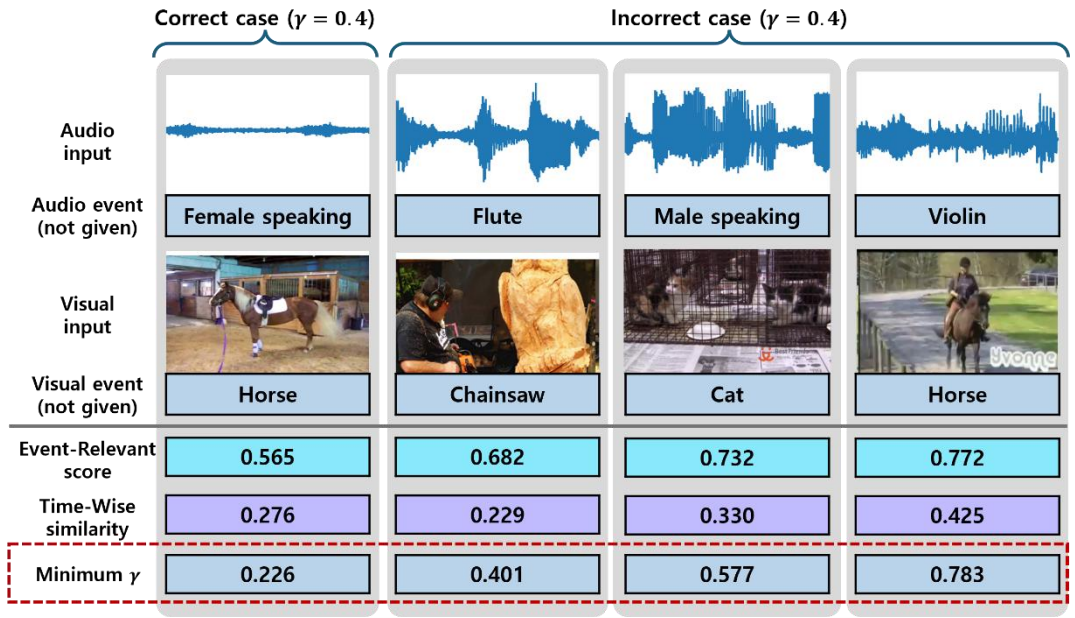


Illustration of examples for failure analysis

Conclusion and Contribution

- This study aims to address the problem of conventional models reaching erroneous predictions, such as incorrectly predicting “Background” instead of “Event,” when inconsistent audio-visual information is clearly present.
- To address this, a GCR method based on a fine-tuned AudioCLIP [7] is proposed, leveraging time-wise similarity between audio and visual inputs to adjust modality significance values when the inconsistent information is clearly presented.
- The experimental results demonstrate that the proposed method outperformed conventional methods on AVE, and future work directions are suggested by analyzing the results through event-relevant prediction and event classification.

Achievement

- Experiences
  - Intern, PIA Space, Seoul, Korea
- Principal Investigator
  - CLIP-based Noise-Robust Audio-visual Event Localization, National Research Foundation of Korea
- Events
  - 2024 The AI Show [AutoML Lab Booth], Seoul, Korea, 2024

Thank you

Reference

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Dec. 5-10, 2016. “SoundNet: Learning sound representations from unlabeled video”. In Proceedings of Advances in Neural Information Processing Systems, Vol. 29. Curran Associates, Inc., Barcelona, Spain, 892–900.

[2] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. “Audio-Visual event localization in unconstrained videos”. In Proceedings of the European conference on computer vision (ECCV), pages 247–263, Munich, Germany, 8-14 Sep. 2018.

[3] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan. “Cross-Modal relation-aware networks for audio-visual event localization”. In Proceedings of the 28th ACM International Conference on Multimedia, pages 3893–3901, Seattle, United States, 12-16 Oct. 2020.

[4] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang. “Positive sample propagation along the audio-visual event line”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8436–8444, Virtual, 19-25 Jun. 2021.

[5] Y. Xia and Z. Zhao. “Cross-Modal background suppression for audio-visual event localization”. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 19989–19998, New Orleans, Louisiana, 21-24 Jun. 2022.

[6] T. Mahmud and D. Marculescu. “AVE-CLIP: AudioCLIP-based multi- window temporal transformer for audio visual event localization”. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 5158–5167, Waikoloa, Hawaii, USA, 4-6 Jan. 2023.

[7] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. May. 22-27, 2022. “AudioCLIP: Extending clip to image, text and audio”. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Singapore, 976–980.