# Listen With Seeing: Cross-Modal Contrastive Learning for Audio-Visual Event Localization

Chao Sun, Min Chen, Chuanbo Zhu, *Graduate Student Member, IEEE*, Sheng Zhang, Ping Lu, and Jincai Chen, *Member, IEEE*

***Abstract*—In real-world physiological and psychological scenarios, there often exists a robust complementary correlation between audio and visual signals. Audio-Visual Event Localization (AVEL) aims to identify segments with Audio-Visual Events (AVEs) that contain both audio and visual tracks in unconstrained videos. Prior studies have predominantly focused on audio-visual cross-modal fusion methods, overlooking the fine-grained exploration of the cross-modal information fusion mechanism. Moreover, due to the inherent heterogeneity of multi-modal data, inevitable new noise is introduced during the audio-visual fusion process. To address these challenges, we propose a novel Cross-modal Contrastive Learning Network (CCLN) for AVEL, comprising a backbone network and a branch network. In the backbone network, drawing inspiration from physiological theories of sensory integration, we elucidate the process of audio-visual information fusion, interaction, and integration from an information-flow perspective. Notably, the Self-constrained Bi-modal Interaction (SBI) module is a bi-modal attention structure integrated with audio-visual fusion information, and through gated processing of the audio-visual correlation matrix, it effectively captures inter-modal correlation. The Foreground Event Enhancement (FEE) module emphasizes the significance of event-level boundaries by elongating the distance between scene events during training through adaptive weights. Furthermore, we introduce weak video-level labels to constrain the cross-modal semantic alignment of audio-visual events and design a weakly supervised cross-modal contrastive learning loss (WCCL Loss) function, which enhances the quality of fusion representation in the dual-branch contrastive learning framework. Extensive experiments conducted on the AVE dataset for both fully supervised and weakly supervised event localization, as well as Cross-Modal Localization (CML) tasks, demonstrate the superior performance of our model compared to state-of-the-art approaches.**

***Index Terms*—Audio-visual event localization, audio-visual information integration, cross-modal contrastive learning.**

Received 18 September 2023; revised 8 May 2024 and 27 June 2024; accepted 22 September 2024. Date of publication 28 January 2025; date of current version 12 May 2025. This work was supported by the National Natural Science Foundation of China under Grant 62272178 and Grant 62276109. The associate editor coordinating the review of this article and approving it for publication was Prof. Zheng-Jun Zha. *(Corresponding Author: Jincai Chen.)*

Chao Sun, Chuanbo Zhu, Sheng Zhang, Ping Lu, and Jincai Chen are with the Wuhan National Laboratory for Optoelectronics, and Key Laboratory of Information Storage System, Engineering Research Center of Data Storage Systems and Technology, Ministry of Education of China, and School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: chaosun@hust.edu.cn; chuanbo_zhu@hust.edu.cn; zhangmonkey@hust.edu.cn; luping06@hust.edu.cn; jcchen@hust.edu.cn).

Min Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China, and also with Pazhou Laboratory, Guangzhou 510640, China (e-mail: minchen@ieee.org).

The code will be available from https://github.com/Supersunn/CCLN.

Digital Object Identifier 10.1109/TMM.2025.3535359

## I. Introduction

In A noisy environment, understanding the speaker's content becomes easier when observing the speaker's facial expressions and body movements in addition to listening to their voice. Similarly, upon hearing a train whistle, individuals instinctively scan their surroundings to locate the source. These examples underscore the cognitive ability of the brain to integrate audio and visual information [1]. According to the Sensory Integration Theory [2], various sensory information inputs (visual, auditory, olfactory, etc.) are transmitted and interact as bio-electrical signals, culminating in integration within the cerebral cortex, thereby facilitating decision-making and consciousness [3].

As multimedia becomes the predominant information medium, the advent of video platforms like YouTube has opened avenues for multimodal tasks in artificial intelligence [4]. Numerous endeavors, spanning lip-reading [5], [6], sound/video event detection [7], [8], [9], [10], sound synthesis [11], emotion recognition [12], [13], and more, aim to endow machines with human-like perception of external stimuli [4]. The complementary relationship between audio and visual cues enriches our understanding of objects and scenes, leading to significant advancements in tasks reflecting audio-visual coordination. These tasks include audio-visual correspondence (AVC) [14], [15], audio-visual instance discrimination (AVID) [16], [17], [18], and audio-visual event localization (AVEL) [19], [20], [21]. AVEL represents an artificial intelligence task centered around the integration of audio-visual information and the localization of audio-visual events. The Audio-Visual Event (AVE) dataset [19], derived from Audioset [22], a large-scale dataset of audio-visual events sourced from YouTube videos, serves as the foundation for AVEL. Each sample in the AVE dataset is an unconstrained video with both audio and visual tracks, encapsulating an audio-visual event (illustrations are presented in Fig. 1).

Physiological and psychological research has elucidated that semantic coherence plays a pivotal role in the integration of multi-sensory input [3], while signal synchronization stands out as a key factor in cross-modal perception integration [4]. As depicted in Fig. 2, incomplete or dysfunctional audio-visual sensory integration processes can lead to advanced audio-visual dysfunction. Indeed, the issue of audio-visual modality misalignment (or inconsistency) pervades real-life unconstrained videos, primarily manifesting in two dimensions. (1) From a visual modality perspective, the susceptibility of two-dimensional

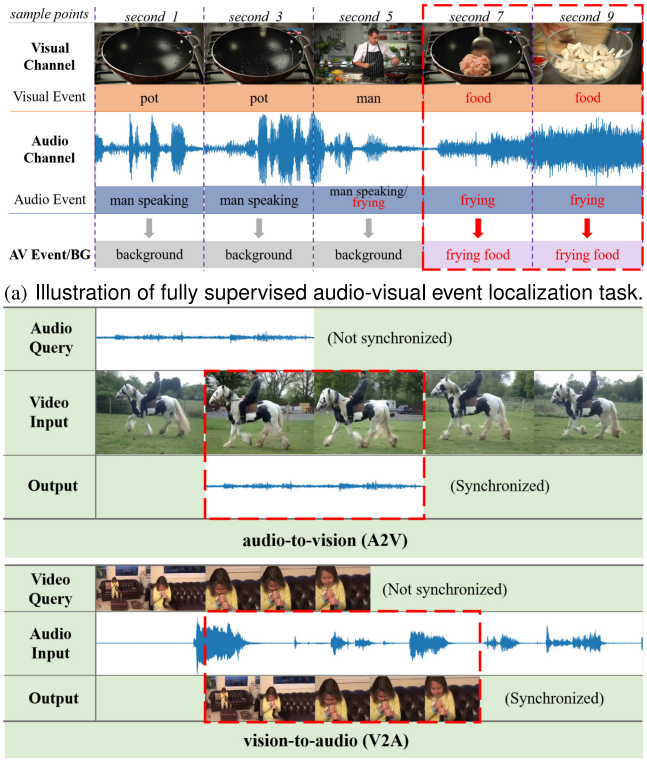



Fig. 1. Illustration of audio-visual event localization tasks. (a) Illustration of fully supervised AVEL. We sample the corresponding audio and visual frames at equal intervals from a 10-second video. The visual frames depict three objects: "pot","man", and "food", while the audio frames capture two sound events: "man speaking" and "frying". An audio-visual event is identified only when the visual object aligns semantically with the sound event, such as "frying food". All other combinations of audio and visual cues are labeled as "background". (b) Illustration of cross-modality AVEL. It aims to query the event boundary of one modality from the corresponding input of another modality. Specifically, visual localization from an audio sequence query is referred to as "vision-to-audio", and audio localization from a video sequence query is referred to as "audio-to-vision".

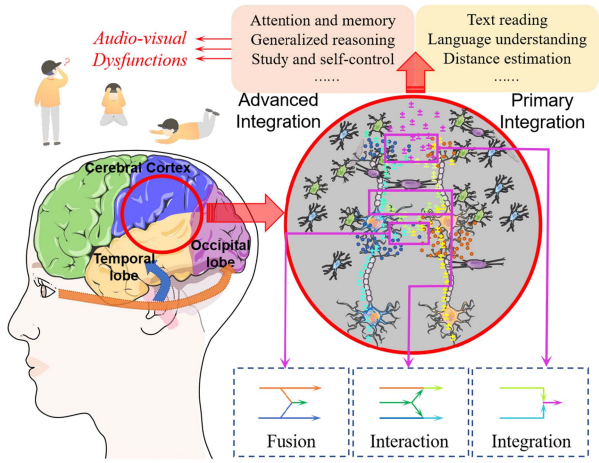



Fig. 2. Illustration of the sensory integration flowchart of the brain. audio and visual signals are transmitted in the form of bio-electricity to the occipital and temporal lobes respectively. Then they are further integrated in the cortex to form primary and advanced integration consciousness. If the audio-visual sensory integration process is incomplete or dysfunctional, it will lead to audio-visual dysfunction. We extract three stages from this information process: fusion, interaction, and integration.

sound signals to noise and the complexity of sound sources render the audio content more uncontrollable. In certain scenarios, the sound producer may not be visible in the video (e.g., voice-over), while in others, multiple sound sources in the environment can introduce interference. (2) From an audio modality perspective, visual scenes tend to harbor more content targets and richer external interference (e.g., exposure, deformation, watermarking, etc.), further complicating the identification of audio content.

Early works address the audio-visual modality misalignment for AVEL using fusion-based frameworks. They tend to focus on intra-modal information fusion methods. In their fusion stages, single-modal features are input into an attention module [19], [23], a Long Short-Term Memory (LSTM) network [24], a Multimodal Factorized Bilinear (MFB) model [25], or a well-designed Transformer module [26]. Subsequently, many works [20], [21], [27], [28], [29], [30], [31], [32] introduce residual lines or self-attention modules to interact cross-modal information, enhancing the model's audio-visual matching capability. To facilitate a fine-grained exploration of the fusion mechanism of audio-visual information, inspired by the sensory integration theory, we attempt to explain the AVEL problem from the perspective of information flow. Hence, we abstract the audio-visual signal processing into three stages (as shown in Fig. 2): merging the original audio-visual signals into a new signal (Fusion-stage), independently incorporating the new signal into the audio or visual signal (Interaction-stage), and aggregating the incorporated audio or visual signals into a decision signal (Integration-stage). Based on the above analysis, we summarize three main issues of AVEL: (1) Pure global fusion is insufficient to better represent audio-visual information, and it is necessary to explore fine-grained intra-modal and inter-modal feature fusion methods for audio-visual pairs; (2) To align the semantic of audio-visual events, the event boundaries in the audio modality are often not prominent, while events in the visual modality are often influenced by changes in camera angles. This directly leads to challenges in capturing the semantic boundaries of audio-visual events; (3) In terms of model structure, models with interaction stages have poor robustness to misaligned audio-visual events. Misaligned audio-visual pairs are more prone to amplifying semantic errors during cross-modal information interaction, which is inevitable in the structure of fusion-based frameworks.

To address these issues, we propose a novel Cross-modal Contrastive Learning Network (CCLN) framework for AVEL, which comprises a backbone network and a branch network. The backbone network is a fusion-based architecture aimed at fully exploiting intra-modal and inter-modal event information. In the backbone network, we decompose the perception integration process of audio-visual events into three stages: fusion, interaction, and integration, delving into the fine-grained investigation of audio-visual fusion mechanisms. The Self-constrained Bi-modal Interaction (SBI) module is specifically designed for the fusion and interaction of audio-visual information. It

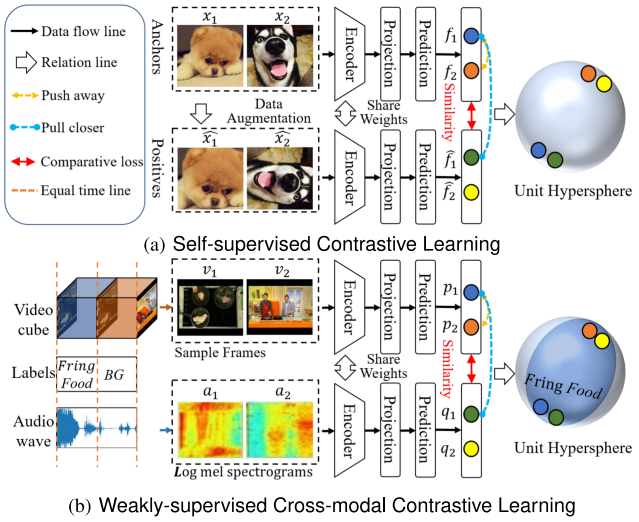



Fig. 3. The different procedures between self-supervised contrastive learning and our proposed weakly-supervised cross-modal contrastive learning. Self-supervised contrastive learning (illustrated in subfigure (a)) contrasts each anchor sample and its augmentation with the remaining negative samples in the batch, which can be seen as a semantic clustering problem. Our weakly-supervised cross-modal contrastive learning (illustrated in subfigure (b)) generates a classification sub-hypersphere guided by weak labels. Inside the sub-hypersphere, all samples of the same class undergo semantic clustering to reduce intra-class distance; outside the sub-hypersphere, all samples of the same class serve as positive samples and contrast with the negative samples in the batch to stretch inter-class distance.

computes audio-visual correlation matrices in a self-attention manner, which are then fused separately into the visual and audio modalities after gate thresholding. The SBI module effectively serves as both a visual-guided audio and an audio-guided video fusion pluggable structure. The fused features capture additional cross-modal details that positively contribute to information interaction. Additionally, we address the event-level audio-visual semantic enhancement, which has been rarely considered in previous models. In the backbone network, we propose the Foreground Event Enhancement (FEE) module, which strengthens event semantic boundaries with adaptive weights while reducing the impact of background event noise. To reduce new noise interference introduced by audio-visual integration and capture weak cues of audio-visual associative semantics, we introduce weak labels to enhance the self-supervised contrastive learning framework, transforming it into a cross-modal feature semantic aggregation problem within the weak label domain (as shown in Fig. 3(b)). Unlike most anchors that consider only single-modal and single-positive factors, our Weakly-supervised Cross-modal Contrastive Learning Loss (WCCL Loss) considers more multi-modal negative and positive factors. These positives come from samples of the same category as the anchor, rather than from anchor data augmentation. The main innovations of our work are:

- We investigate and elucidate the audio-visual fusion mechanism of the AVEL task from the perspective of information flow, dividing the audio-visual fusion process into three stages: fusion, interaction, and integration. This provides a novel approach to uncovering hidden cross-modal complementary information and enhancing the coupling of audio-visual pairs.
- We incorporate pluggable SBI and FEE modules in the fusion-based backbone network. The SBI module utilizes a bi-modal attention structure integrated with audio-visual fusion information to effectively capture inter-modal correlations, while the FEE module emphasizes the significance of event-level boundaries, better capturing the weak semantic boundaries of audio-visual events.
- We propose a two-branch contrastive learning framework for AVEL for the first time to reduce the new noise introduced during audio-visual fusion. To solve the semantic alignment problem of cross-modal contrastive learning, we introduce weak labels to constrain the audio-visual event semantics, so that the model can obtain better cross-modal semantic features.
- Experimental results on the extensively utilized AVE dataset demonstrate that our proposed model surpasses the state-of-the-art methods for both fully supervised and weakly supervised event localization, as well as cross-modal localization tasks.

## II. Related Work

### A. Audio/video Anomaly Event Detection

Traditional audio/video anomaly event detection entails a binary classification task aimed at discerning the presence of audio/visual anomaly events throughout the entire audio/video scene. Treating an audio/video abnormal segment as an event broadens the purview of audio/video anomaly detection to encompass single-modal (i.e., audio or visual) event localization. Anomalies (i.e., anomalous events) typically manifest for brief durations in real-world scenarios. Therefore, prior endeavors have endeavored to establish normal patterns using various statistical models and classify segments diverging from these patterns as abnormal events. Commonly employed methods for identifying outliers as anomalies include Hidden Markov Model (HMM) [33], Gaussian process modeling [34], sparse reconstruction methods [35], and clustering-based approaches [36]. However, these methods may not effectively capture audio/video time-series cues.

With the significant advancements in deep learning, some researchers have turned to generative models for constructing normal behavior patterns, including Generative Adversarial Networks (GAN) [37] and Autoencoders [8], [9]. Additionally, the Seg2Seg framework [38] has been widely applied to leverage the temporal continuity of audio/video data. Sultani et al. [39] propose a multi-instance weakly supervised framework for predicting visual normal/abnormal behavior, departing from the sole modeling of normal behavior. This approach has yielded promising results and has been further explored [38], [40]. In recent years, self-supervised learning models [41], [42], relying on data augmentation, have also gained considerable attention in the anomaly detection domain. These models generate spatial and temporal pseudo-abnormal data for self-supervised training alongside normal data. While these methods typically focus on either audio or visual signals, we concurrently consider two

types of heterogeneous data from different modalities. Similar to single-modal event detection methods, we extract event features from segments.

### B. Audio-Visual Representation Learning

Audio-visual representation learning endeavors to obtain high-quality joint representations of audio-visual pairs, requiring the comprehensive utilization of complementary information within and across audio-visual modalities. Early methods for audio-visual representation were limited by computational resources and mainly relied on mathematical and statistical approaches. However, with the rise of deep learning, audio-visual representation has undergone significant evolution, embracing fusion strategies rooted in both supervised and unsupervised learning paradigms.

In most models, the prevailing paradigm operates under supervised learning. These models typically utilize dual branches to extract and process features from audio and visual channels, subsequently employing a fusion module to integrate these features. They are trained using actual audio-visual labels as supervised signals [43], [44], [45], [46]. For instance, Min et al. [43] devise four distinct families of objective A/V quality prediction models employing diverse multi-modal fusion strategies. Xue et al. [44] introduce a co-attention model to supplant direct multi-modal fusion, leveraging spatial and semantic correlations between audio and visual features. Delving deeper into audio-visual relationships with attention-based networks, Liu et al. [45] employ a dense modality interaction network integrating two innovative modules to harness audio-visual information. Conversely, significant strides have been made in audio-visual representation learning through unsupervised (including self-supervised) strategies. These approaches hinge on semantic alignment achieved via contrastive learning losses. For example, Owens et al. [47] endeavor to learn joint cross-modal representations, considering sound and corresponding visual images as supervisory signals in an unsupervised manner. Zheng et al. [48] seek to generate modality-independent representations for each individual in each modality via adversarial learning, concurrently learning robust similarity for cross-modal matching through metric learning. In this study, the fusion-based backbone network we propose is rooted in the analysis of information-flow transfer stages, with pluggable SBI and FEE components designed to enhance the internal correlation of audio-visual representation.

### C. Audio-Visual Event Localization

Audio-visual event localization aims to identify audio-visual events of interest within unconstrained, long video sequences and predict the category to which these events belong. Specifically, the AVE task involves discovering event-matching video segments within video sequences that contain both audio and visual events (with the background considered as one event). Subsequently, predictions are made regarding the categories of audio-visual events, either at the segment level or video level. Early models primarily focused on fusion methods for audio-visual signals, including early fusion of audio-visual features and late fusion into predictions (i.e., integration). Tian et al. [19] were the first to propose the AVE task and demonstrate the effectiveness of audio-guided visual attention (AGVA), which has become an important component of most subsequent models. Lin et al. [24] leverage the temporal-dependence properties of LSTM to concatenate audio-visual single-modal features for predictions. Ramaswamy et al. [25] incorporate a bi-linear model to integrate the extracted audio-visual features. Lin et al. [26] construct a complex audio-visual Transformer structure with an AGVA module to capture the relationship information between audio and visual features. Xu et al. [23] design an attention structure and integrate predictions using matrix dot product.

To further explore the complementary information of audio and visual signals, and to reduce the noise generated during the process of audio-visual fusion, many researchers have focused on the multi-modal information interaction and integration processes to construct more sophisticated networks. Wu et al. [27] introduced a self-attention module to integrate intra-modal information dependencies and incorporated residual connections for basic inter-modal information interaction. Xuan et al. [29] utilized self-attention, adaptive attention, and LSTM modules to create a network structure for audio-visual information interaction through residual connections. Ramaswamy et al. [28] leveraged bilinear methods to achieve a more complex audio-visual information interaction process, considering additional interaction information during prediction generation. Zhou et al. [30] employed a threshold to filter out strongly related event segments during the information interaction stage but also discarded potentially valuable relevant information. Xia et al. [20] improved the design of the information integration stage and used an attention-based approach to suppress noise at both the temporal and event levels. Wang et al. [21], [32] emphasized the importance of event boundaries, advocating for finer-grained modulation of segment-level semantics and event-level relationships following the fusion and interaction stages.

In our model, we thoroughly explore the fusion mechanism of audio-visual information, refine the structural design, and propose a cross-modal contrastive learning paradigm to reduce the new noise generated by audio-visual fusion, conducting in-depth research on various structural aspects.

### D. Audio-Visual Contrastive Learning

In recent years, self-supervised contrastive learning models have witnessed significant advancements across various domains [49]. Typically, the input to a self-supervised contrastive learning model involves utilizing a positive pair, chosen through co-occurrence [49], [50] or data augmentation [51], for each anchor sample. This selection method is often based on limited prior knowledge, such as frames from different videos or patches from distinct images, aimed at enhancing the model's accuracy. Tian et al. [50] were the pioneers in exploring the multi-view coding (CMC) technique within a contrastive learning framework, intending to encode various data views (e.g., brightness, optical flow) from the same image sample. However, these data

views primarily pertain to the visual modality and inherently possess pre-existing semantic features. An intriguing concept in self-supervised contrastive learning involves substituting the positive pairs of the same modality with embeddings from different modalities, such as audio and video [16], [17].

Typically, contrastive loss is employed during the training stage to minimize the distance between representations in the last layer of a deep network. However, for multi-modal tasks, the heterogeneity across modalities significantly reduces confidence in representations based solely on feature similarity. Furthermore, self-supervised contrastive learning based on a single modality often requires heavy data augmentations to generate diverse views. To address the challenges posed by self-supervised contrastive learning in the presence of multi-modal heterogeneity, an effective approach is to introduce labels to narrow or push the distance between multi-modal samples. Kamnitsas et al. [52] introduce a novel regularization method and apply it to joint training classification heads with contrastive embeddings. Subsequently, Khosla et al. [53] increase data augmentation, normalize the contrastive embeddings, and propose a supervised contrastive loss (SupCon loss), which achieves remarkable results across numerous pretext tasks. Inspired by the SupCon loss, we introduce weak labels to mitigate noise introduced by modality heterogeneity in cross-modal representation learning. However, our approach differs from the SupCon loss in that we deal with multi-modal data with weak labels at the segment or video level, while the SupCon loss is applied to single-modality (image) data with fine labels. Additionally, the SupCon loss establishes a multi-positive pattern for each anchor, whereas, in the AVEL task, we must consider not only the label for each feature but also the alignment of multi-modal data within the same sample.

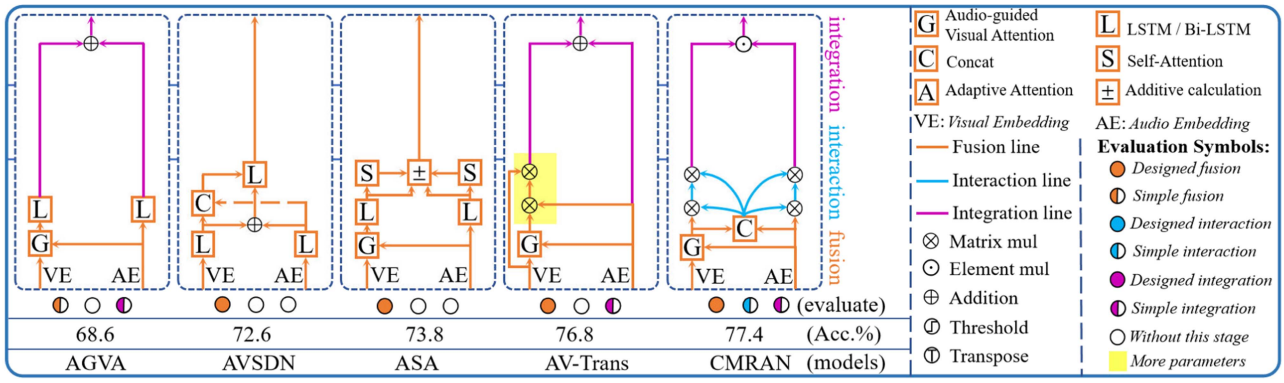
## III. Motivation

Traditionally, the fusion-based framework for AVEL has predominantly focused on integrating audio and visual information flows, beginning from low-dimensional data and progressing to higher-dimensional semantic information. Analogous to the way human auditory and visual signals are transmitted and integrated into the cerebral cortex to facilitate advanced audio-visual functions, the design of AVEL frameworks should meticulously consider audio-visual information processing. Hence, we have delineated three stages of audio-visual information processing: fusion (Fusion-stage), interaction (Interaction-stage), and integration (Integration-stage), mirroring the mechanism by which humans process audio-visual signals. However, are all three stages indispensable for AVEL? What challenges may arise from the fusion-based model paradigm, and how have they been addressed? We aim to distill the essential elements of mainstream AVEL frameworks to analyze the mechanism of audio-visual information processing, as illustrated in Fig. 4.

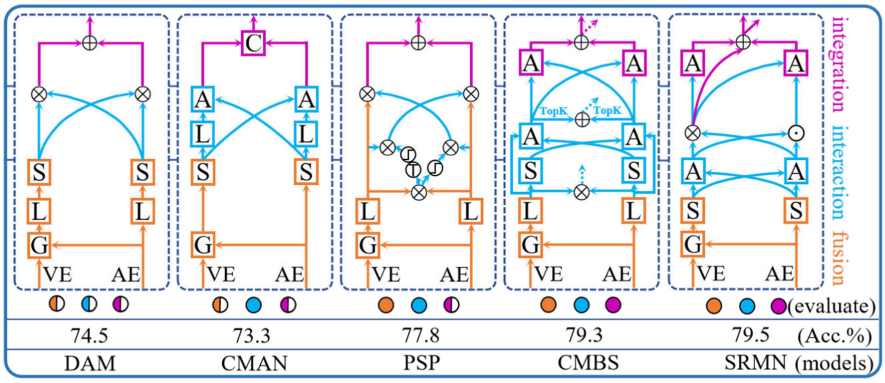
*Is the Fusion-stage necessary?* Almost none of the model architectures depicted in the first row of Fig. 4 incorporate an information Interaction-stage; instead, they solely focus on information fusion. The AGVA model [19] demonstrates the efficacy of audio-guided visual attention for the first time, with its fusion architecture serving as the basis for most subsequent models. In contrast to AGVA, the AVSDN model [24] concatenates LSTM-encoded audio-visual features into a global LSTM for fusion. The ASA model [25] introduces a self-attention-like module based on AGVA and computes cross-modal fusion information using addition. The AV-Trans model [26] integrates a fusion module with a larger parameter scale, aiming to construct more refined temporal/spatial fusion methods using the Transformer [54] architecture. The CMRAN model [23] adopts multiple attention modules, introducing temporal information to enhance AGVA and fusing single-modal information with two self-attention structures, respectively. Comparative analysis reveals that both AVSDN and ASA outperform the baseline AGVA, suggesting that the designed Fusion-stage effectively enhances model performance. Furthermore, the performance results of AV-Trans and CMRAN indicate that a sophisticated fusion module with a large parameter scale can significantly leverage the complementarity of cross-modal information.

*Is the Interaction-stage necessary?* The model architectures depicted in Fig. 4(b) encompass, to varying extents, the three stages of information processing (fusion, interaction, and integration). However, they often lack emphasis on the design of the information integration stage, frequently implementing it through simplistic operations such as addition, multiplication, or concatenation. For instance, the DAM model [27] and CMAN model [29] introduce straightforward residual lines to facilitate cross-modal information interaction. The AVIN model [28] focuses on designing a complex information interaction module but overlooks early information fusion, achieving results comparable to AV-Trans despite having smaller-scale parameters. Conversely, the PSP model [30] adopts a more comprehensive architecture design, incorporating an interaction module that leverages earlier fusion information. Following the fusion stage in the CMBS model [20], a more intricate interaction module is established, and improvements are made to the integration stage, resulting in enhanced model performance. Notably, the frameworks presented in Fig. 4(b) consistently achieve improved model performance with the addition of an Interaction-stage, even if they have a simpler Fusion-stage (e.g., AVIN) compared to the baseline AGVA. Particularly, PSP (with an Interaction-stage) outperforms AV-Trans (without an Interaction-stage but with a larger parameter scale) and CMRAN (with a more complex fusion module but a weak Interaction-stage), highlighting the critical role of the Interaction-stage. The overall performance results of these models further affirm that well-designed information interaction significantly benefits AVEL.

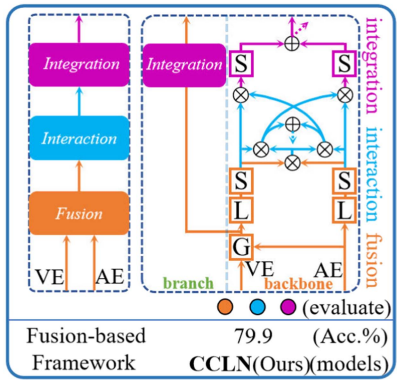
*Is the Integration-stage necessary?* Among the models illustrated in Fig. 4(b), SRMN stands out for its comprehensive design of audio-visual information processing, encompassing fusion, interaction, and integration stages, and it has demonstrated superior performance compared to other models. However, attributing the superior performance of SRMN solely to the design of the information processing stages or to the influence of the event proposal modulation strategy remains challenging. Similarly, previous studies either overlook or incompletely address the design of the information Integration-stage, making it difficult to draw definitive conclusions. Upon




(a) Fusion-based Frameworks (mainly focus on the fusion-stage)




(b) Fusion-based Frameworks (with fusion, interaction, integration)




(c) Two model frameworks

Fig. 4. Comparison of audio-visual information processing frameworks for AVEL. The networks in subfigure (a) only highlight the role of the Fusion-stage and give little consideration to information interaction and integration. The networks in subfigure (b) contain fusion, interaction, and integration stages, most of which do not have a complete structural design. These networks can be abstracted into the fusion-based framework (shown on the left of subfigure (c)). To reduce the noise generated by information fusion, we propose a cross-modal contrastive learning network framework (shown on the right of subfigure (c)), which is the first dual-branch architecture framework for AVEL. The performance results reported in these figures are all under the fully supervised setting.

examining Fig. 4(a), we observe that despite the complex fusion-stage designs in AVSDN and ASA (which lack an integration-stage), their performance is notably inferior (by approximately 3% on average) compared to AV-Trans and CMRAN (both of which incorporate a simple integration-stage). This observation suggests that the integration-stage may indeed play a beneficial role in audio-visual cross-modal information processing. Nonetheless, specific ablation experiments are necessary to validate this hypothesis.

*Assumption:* Based on the above analysis, we can outline a general audio-visual information processing paradigm for AVEL, as depicted on the left side of Fig. 4(c). Additionally, insights drawn from previous frameworks (illustrated in (a) and (b) of Fig. 4) provide valuable inspirations for designing our framework: (1) Attention mechanism has been proven to be beneficial in the information Fusion-stage, and its reasonable use can achieve good model performance (refer to the results of CMRAN); (2) Using LSTM to enhance the dependency of intra-modal information is a good choice; (3) The larger the model parameter scale, the greater the performance advantage (refer to the results of AV-Trans), but it also introduces more noise; (4) Information interaction should be considered, as demonstrated by high-performance models such as the SRMN; (5) The contribution of the information integration-stage requires further experimental exploration, and the treatment of new noise in the audio-visual information processing should also be considered.

Based on the considerations outlined above, we introduce a novel AVEL model paradigm (depicted on the right side of Fig. 4(c)). This paradigm incorporates a backbone network comprising three stages: information fusion, interaction, and integration, which collectively constitute the Cross-Modal Contrastive Learning Network (CCLN). The detailed methodology of our entire model is delineated in Section IV, while Section V-B presents the fundamental aspects of each stage through experimental analyses. In Fig. 5, we illustrate four specific branch structures of the CCLN, and a comparative investigation is further conducted in Section V-B.

## IV. Methodology

Analogous to the human brain's integration of multi-sensory information, we conceptualize audio-visual event localization as a process involving the fusion, interaction, and integration of audio-visual pairs to predict and categorize events at either the segment or video level.

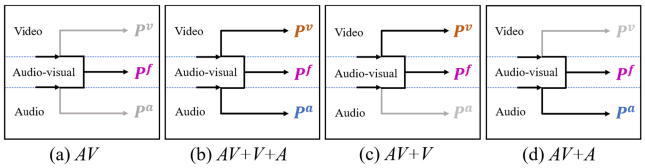



Fig. 5. The four contrastive branch structures of CCLN are illustrated. In the figure, (a) mode denotes only the audio-visual fusion branch (AV), (b) mode represents contrastive learning between the audio-visual fusion branch and the visual branch, and the audio branch (AV+V+A), (c) mode indicates contrastive learning between the audio-visual fusion branch and the visual branch only (AV+V), and (d) mode signifies contrastive learning between the audio-visual fusion branch and the audio branch only (AV+A).

### A. Notations and Problem Statement

The goal of AVEL is to identify synchronously related audio-visual pairs, content-matching audio and visual events, from an unconstrained video. Suppose a video sequence containing audio and visual modalities is $\{S_t^v, S_t^a\}_{t=1}^T$. $S_t^v$ and $S_t^a$ represent the visual and audio channels of the video sequence respectively. The video sequence is divided into non-overlapping $T$ segments at equal intervals, and the sampling interval is taken as one second in this paper. Consistent with the baseline [19], AVEL explores learning under both fully supervised and weakly supervised learning settings. The model of cross-modal localization task, however, is learned under a fully supervised learning setting.

*AVEL under Fully Supervised Learning:* In the fully supervised learning setting, each audio-visual video segment is assigned an event category label, and the model is trained to predict the event category at the segment level. $y_t^f = \{y_t^{fk} | y_t^{fk} \in \{0,1\}, \sum_{k=1}^{C+1} y_t^{fk} = 1\}$ represents the label of the $t^{th}$ segment, where $t \in \{1, 2, \ldots, T\}$. Note that we treat the background as an independent event category, where $C$ is the number of label categories of a dataset, so the total number of event categories is $C+1$. Thus, define $Y^{full} = \{y_1^f, y_2^f, \ldots, y_T^f\} \in \mathbb{R}^{T\times(C+1)}$ as the label of the entire video sequence. The prediction score of the $t^{th}$ audio-visual pair in the fully supervised training setting can be used to judge the event category of the segment.

*AVEL under Weakly Supervised Learning:* In the weakly supervised learning setting, we can only get the video-level labels. $Y^{weak} = \{y^{wk} | y^{wk} \in \{0,1\}, \sum_{k=1}^{C+1} y^{wk} = 1\} \in \mathbb{R}^{(C+1)}$ is denoted the label of an entire video. This setting is more suitable for real-world general situations where fine annotations are not readily available but poses a higher challenge to the robustness of the model. In this paper, we perform linear transformation and average pooling on $Y^{full}$ to obtain the video-level labels.

*Cross-modality Localization Task of AVEL:* The cross-modal localization task of AVEL aims to determine the boundaries corresponding to audio-visual events and is conducted under fully supervised training at the segment level. In either the audio or video modality, each segment's label is binary, denoted as 0 or 1. Consequently, the label for the entire video segment can be represented as $Y^{cml} = \{(y_1, y_2, \ldots, y_T) | y_t \in \{0,1\}\} \in \mathbb{R}^{T\times 1}$. It is noteworthy that the model for the CML task does not necessitate the design of an audio-visual fusion structure or a classification head; its ultimate objective is to compute the distance between predicted segments in the query modality and given modality event segments.

### B. Overall Model Structure

Fig. 6 illustrates the overall structure of our CCLN model, which primarily comprises a backbone network and a branch network (visual branch). The backbone network is composed of five key modules: (1) Feature Embedding: This module separates audio and visual frames from an unconstrained long video, transforms and encodes them into high-dimensional feature vectors; (2) Fusion-Stage: Using audio embeddings to guide visual embeddings through a co-attention strategy, this module achieves cross-modal fusion and enhances inter-modal temporal dependence; (3) Interaction-Stage (SBI module): In this stage, audio and visual information interact, with the degree of interaction controlled by adaptive parameters;(4) Integration-Stage (FEE module): Based on the integration of inter-modal information, this module employs the FEE branch to reduce background event noise; (5) Classification: In this module, decision information is utilized to train a classifier under the constraint of the loss functions, which makes predictions based on the training results.

The preprocessed audio and video segments are fed into the feature embedding module, where a deep convolutional network abstracts them into high-dimensional semantic features respectively. In the Fusion-stage, the audio-visual features are aligned by a spatial attention structure (i.e. audio-guided visual attention model [19]), and then a Bi-LSTM component [24] and a self-attention component is applied to the audio and visual modalities separately to strengthen the long and short term dependence of the intra-modal information. Subsequently, the encoded feature $v^F$ and $a^F$, corresponding to audio and visual, are sent to the SBI module for refined interaction of audio-visual signals. In the Interaction-stage, a bi-attention structure is designed to filter and fuse the correlation information between audio-visual modalities, and the correlation information as well as the cross-modal residual information are used to guide the generation of visual feature $v^I$ and audio feature $a^I$. Then, in the Integration-stage, based on integrating the intra-modal information of audio and visual modality, the adaptive weight branch of the FEE module is introduced to improve the attention of foreground events. Finally, in the classification module, the integrated information $v^Z$ and $a^Z$ are linearly transformed to generate the audio-visual event predictions. Note that $v^F$, $a^F$, $v^I$, $a^I$, $v^Z$ and $a^Z$ are all vectors of dimension $\mathbb{R}^{B\times T\times d_v}$.

### C. Feature Embedding

The task of the feature embedding module is to extract and encode abstract audio and visual representations and unify them into feature vectors of the same dimension. First, the raw audio and visual channels are separated from the unconstrained video containing audio and visual pairs. The visual channel samples $T$ cubes $S^v$ at the same interval, and the audio Mel-spectrum $S^a$ is also sampled at equal intervals into $T$ segments after the Mel-scale filter banks. Each image cube or audio Mel-spectrogram

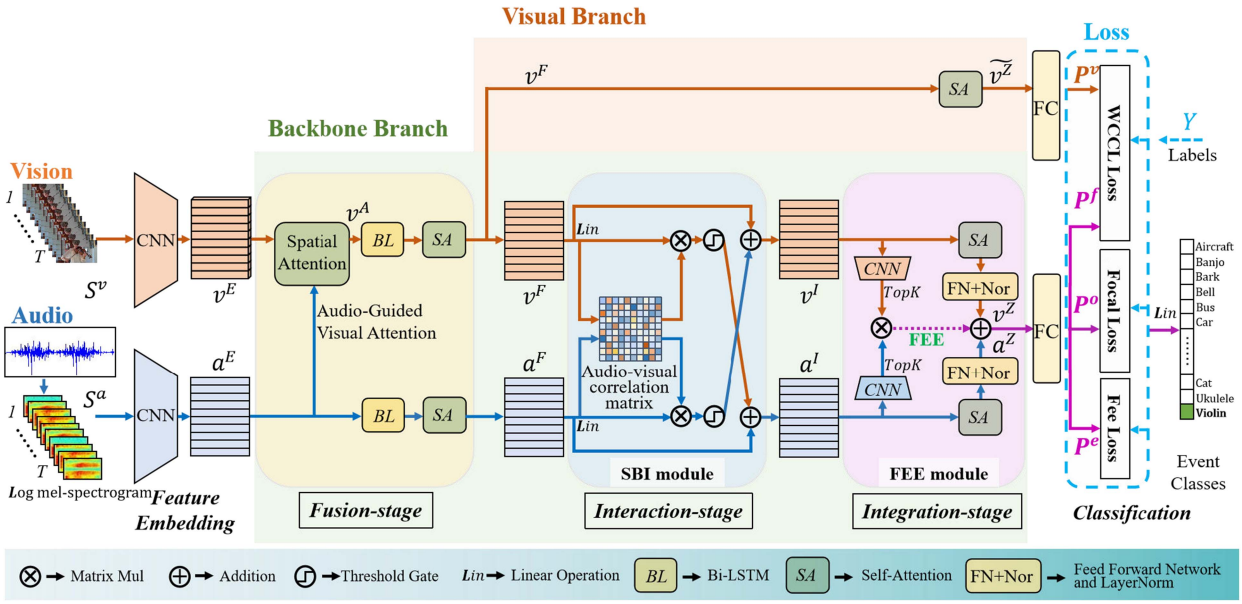



Fig. 6. The overall architecture of our CCLN model. Preprocessed audio and visual raw data are input into pre-trained deep convolutional networks to extract single-modal high-dimensional features, which are then fed into the CCLN dual-branch network. In the fusion-based backbone branch, audio and visual features undergo processing through three stages: fusion, interaction, and integration. In the fusion stage, we employ an audio-guided visual attention module and a Bi-LSTM module for intra-modal information enhancement and initial inter-modal fusion. In the interaction stage, we design the SBI module, where the correlation matrix after audio-visual fusion is used to construct cross-modal attention, which is then fused into audio and visual features after the threshold gate. In the integration stage, the FEE module further filters and integrates to generate event-level predictions, enhancing the boundaries of audio-visual events. The CCLN model utilizes the visual branch after the fusion stage as the contrastive branch.

segment is then fed into an independent pre-trained deep convolutional neural network (CNN) to extract high-dimensional feature $v^E \in \mathbb{R}^{T\times H\times W\times d_{fv}}$ or $a^E \in \mathbb{R}^{T\times d_{fa}}$. $H$ and $W$ are the height and width of the video frame respectively, and the dimensions of $d_{fv}$ and $d_{fa}$ are not equal here.

### D. Backbone Network

Analogous to the way audio and visual signals are analyzed and transmitted in human brain regions related to audio-visual integration, the backbone branch network serves the purpose of filtering and fusing intra and inter-modal signals in our model. The backbone network mainly consists of three audio-visual information processing modules: the Fusion-stage, the Interaction-stage, and the Integration-stage.

*Fusion-stage:* The Fusion-stage is used to effectively obtain the early audio-visual fusion information and enhance intra-modal information dependence. The audio-guided visual attention (AGVA) mechanism [19] has fully demonstrated that it can adaptively find the corresponding audio object or visual activity from the visual modality of each video segment. Therefore, we employ this spatial attention approach to compensate for audio-visual information in visual features. Then, we adopt Bi-directional Long Short-Term Memory (Bi-LSTM) to establish long short-term dependence information for audio or visual features along the time direction, and the visual representation $v^F \in \mathbb{R}^{T\times d_v}$ and audio representation $a^F \in \mathbb{R}^{T\times d_a}$ are obtained after a self-attention structure. In this process, the dimension $d_v$ of audio features and dimension $d_a$ of visual features are already equal after linear transformation. The whole process can be recorded as:

$$v^A = AGVA\left(v^E, a^E\right) \tag{1}$$

$$v^F = Sa\left(Bl\left(v^A\right)\right) \tag{2}$$

$$a^F = Sa\left(Bl\left(a^E\right)\right) \tag{3}$$

where $Bl(\cdot)$ and $Sa(\cdot)$ represent Bi-LSTM and self-attention operations, respectively.

*Interaction-stage (SBI Module):* Although the use of a more complex Fusion-stage design [26] can improve the performance of the model, it also introduces unexplained and uncancelable noise, which limits the final performance of the model. At the same time, the computational resource consumption caused by a large number of parameters is not necessary for the model performance, and we can exceed its performance by optimizing the information flow. PSP [30] clearly simulates the processing of intra and inter-modal information flows, attempting to find strongly correlated audio-visual pair through a one-hot encoded correlation matrix, but the hard threshold also inadvertently loses richer multi-modal interaction information. CMBS [20] designs a more complete information interaction structure, but the redundancy in the structure often introduces new fusion noise. In the Interaction-stage, we design the SBI module to exchange audio-visual information through a bi-attention structure and apply cross-residual information to compensate for missing information.

The detailed structure of our designed cross-modal information interaction component, the SBI module, is shown in Fig. 7. Among feature representation methods, the attention mechanism

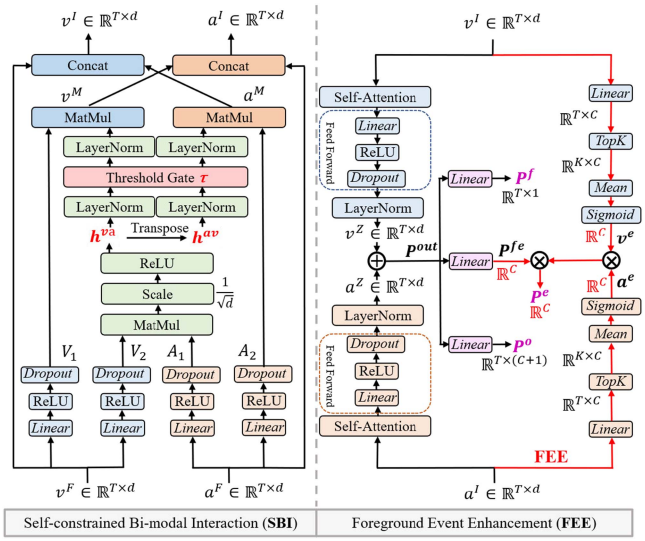



Fig. 7. The architecture of the SBI module (left) and FEE module (right). SBI is a size-invariant pluggable component with dual input and dual output, used for the interaction of audio-visual information under parameter constraints in the interaction stage, effectively balancing cross-modal complementary information and new noise information. In the FEE module, we design a weight-adjustable branch containing only foreground events to capture event boundaries at the event level, reducing noise from background events. Note that the data dimensions in the figure correspond to the fully supervised AVEL.

is superior to the recurrent and convolutional layers in terms of computational complexity and path length between long-term dependencies. The output vectors $v^F$ and $a^F$ of the fusion-stage module are linearly transformed as $V_1$, $V_2$ and $A_1$, $A_2$, respectively. They are all of the same dimension $\mathbb{R}^{T\times d}$, and $d$ is equal to $d_a$. First, we choose Scaled Dot-Product Attention (SDPAtt) to calculate the fusion feature $h^{va}$ of $V_2$ and $A_1$, since it can be implemented using highly optimized matrix multiplication code to achieve much faster and more space-efficient performance. And then the audio-visual correlation matrix $h^{va}$ is transposed to obtain the $h^{av}$ matrix. Subsequently, $h^{va}$ and $h^{av}$ are filtered and normalized with hyperparameter $\tau$ and dot multiplied with $V_1$ and $A_2$ to generate the fused visual feature $v^M$ and audio feature $a^M$. This process is represented as:

$$V_1, V_2, A_1, A_2 = lin\left(v^F W_1, v^F W_2, a^F W_3, a^F W_4\right) \tag{4}$$

$$h^{va} = \sigma\left(\frac{V_2 \otimes {A_1}^T}{\sqrt{d}}\right) \tag{5}$$

$$h^{av} = T(h_{va}) \tag{6}$$

$$v^M = V_1 \otimes h^{va}(\tau) \tag{7}$$

$$a^M = A_2 \otimes h^{av}(\tau) \tag{8}$$

where $W_1$, $W_2$, $W_3$ and $W_4$ are linear transform weights, all with dimension $\mathbb{R}^{d\times d}$. $lin(\cdot)$ denotes the linear transformation, $\sigma(\cdot)$ denotes the Softmax function, $\otimes$ is the dot product, and $T(\cdot)$ denotes the matrix transpose operation.

The SBI module aligns the relevant salient information of the audio and visual features through the fusion strategy, and we adopt the information interaction strategy to further reduce the background noise of the audio-visual pairs. Just as the fusion and interaction of different modality information in the human brain are carried out simultaneously, the interaction process and the fusion process in our model are also carried out simultaneously, which is realized by the residual line. The encoded initial features are cross-added with the fused features $v^M$ and $a^M$ to obtain the interactive features $v^I$ and $a^I$. The SBI module can be regarded as a dimension-invariant dual-input-output pluggable device for bi-modal interaction. This process is written as,

$$v^I = v^F \oplus a^M \tag{9}$$

$$a^I = a^F \oplus v^M \tag{10}$$

*Integration-stage (FEE Module):* Previous audio-visual integration strategies have mostly relied on simple dot products and element-wise multiplication, which can easily result in information loss and introduce noise. Therefore, we employ self-attention structures and Layer normalization to further integrate single-modal information. Additionally, previous methods mostly generate segment-level classification heads, utilizing segment-level or video-level labels for supervised learning. This strategy overlooks the learning of event-level knowledge, making it difficult to distinguish boundaries between audio-visual events. In our FEE module, we construct event-level classification heads, enabling the model to better localize event boundaries.

In the Integration-stage, we integrate the interactive audio-visual information to form the prediction for classification decisions. Firstly, $v^I$ and $a^I$ are pushed into a self-attention component to enhance intra-modal information integration of audio and visual modalities. The output results are then processed by a simple position-wise fully connected feed-forward layer and a normalization layer (i.e. Fl + Ln layer) in turn to produce the $v^Z$ and $a^Z$ features. Finally, the output is simply averaged to get a preliminary decision value of $P^{out}$, and $P^{out}$ is transformed by different linear transformations to get three different predictions $P^f$, $P^{fe}$ and $P^o$. The process is summarized as follows,

$$v^Z = Ln\left(Fl\left(Sa\left(v^I\right)\right)\right) \tag{11}$$

$$a^Z = Ln\left(Fl\left(Sa\left(a^I\right)\right)\right) \tag{12}$$

$$P^{out} = \frac{1}{2}\left(v^Z \oplus a^Z\right) \tag{13}$$

$$P^o = lin\left(P^{out} \otimes W_5\right) \tag{14}$$

$$P^f = lin\left(P^{out} \otimes W_6\right) \tag{15}$$

$$P^{fe} = Max\left(lin\left(P^{out} \otimes W_7\right)\right) \tag{16}$$

where $W_5 \in \mathbb{R}^{d\times(C+1)}$, $W_6 \in \mathbb{R}^{d\times 1}$ and $W_7 \in \mathbb{R}^{d\times C}$ are linear transform weights. The dimensions of $v^Z$, $a^Z$ and $P^{out}$ are both $\mathbb{R}^{T\times d}$. $Fl(\cdot)$ and $Ln(\cdot)$ denote the feed-forward layer and normalization layer.

In particular, we calculate prediction labels for CML tasks from formulas (11) and (12) :

$$P^o_{cml} = sqrt\left(Max\left(\left(v^Z - a^Z\right)^2, 0\right)\right) \tag{17}$$

where $sqrt(\cdot)$ means to find the square root, and $P^o_{cml} \in \mathbb{R}^{T\times 1}$

Due to the high uncertainty of the background event, the FEE module is used to widen the distance between foreground events and background even to reduce background noise. The $C$ dimension of $P^{fe}$ is the category number of audio-visual events in a dataset, that is, the number of foreground events relative to background events. We design another weight branch to calculate the final foreground event prediction. $v^I$ and $a^I$ after linear transformation screen out the first $K$ values through TopK function. The average values of the results are calculated and transformed into $v^e \in \mathbb{R}^C$ and $a^e \in \mathbb{R}^C$ after the activation function. After multiplying $v^e$ and $a^e$, multiply with $P^{fe}$ to get foreground event prediction $P^e \in \mathbb{R}^C$. This process can be expressed as,

$$v^e = \rho\left(Mean\left(TopK\left(lin\left(v^I\right)\right)\right)\right) \tag{18}$$

$$a^e = \rho\left(Mean\left(TopK\left(lin\left(a^I\right)\right)\right)\right) \tag{19}$$

$$P^e = P^{fe} \otimes \left(v^e \otimes a^e\right) \tag{20}$$

where $TopK(\cdot)$ and $Mean(\cdot)$ represent sorting and averaging operations, respectively. $\rho$ is the Sigmoid activation function.

### E. Classification

Prediction $P^o$ is processed differently under different training settings. We defined $W_5$ as $W_{f5}$ (under full supervised settings) and $W_{w5}$ (under weak supervised settings) respectively. This transformation can be described as,

$$P_{full} = lin\left(P^{out} \otimes W_{f5}\right) \tag{21}$$

$$P_w = lin\left(P^{out} \otimes W_{w5}\right) \tag{22}$$

where the dimension of weight $W_{f5}$ and $W_{w5}$ are $\mathbb{R}^{d\times(C+1)}$. So the predictions $P_{full}$ and $P_w$ have the same dimension $\mathbb{R}^{T\times(C+1)}$. In the fully supervised setting, $P_{full} = \{p_1^f, p_2^f, \ldots, p_T^f\} \in \mathbb{R}^{T\times(C+1)}$ is the final prediction score and the $t^{th}$ prediction is represented as $p_t^f = \{p_t^{fk} | p_t^{fk} \in \{0,1\}, \sum_{k=1}^{C+1} p_t^{fk} = 1, t = 1, 2, \ldots, T\}$.

We add a weighted branch in the weakly supervised setting to improve the correlation of captured video-level synchronous audio-visual pairs.

$$\Theta = \rho\left(R\left(lin\left(P_w \otimes W_b\right)\right)\right) \tag{23}$$

$$P_{weak} = P_w \otimes \Theta \tag{24}$$

where $W_b \in \mathbb{R}^{(C+1)\times 1}$ is learnable parameters in the linear layers. $R$ denotes the ReLU activation function. $\Theta \in \mathbb{R}^{T\times 1}$ is the weight vector we get from the linear weight branch. Thus, video-level prediction under weakly supervised setting can be expressed as $P_{weak} = \{p^{wk} | p^{wk} \in \{0,1\}, \sum_{k=1}^{C+1} p^{wk} = 1\} \in \mathbb{R}^{(C+1)}$.

In addition, we take the features after the fusion stage as the input of the branch network and get the prediction features $P^v$ (or $P^a$) after processing by the self-attention component and linear layer. $P^f$, $P^v$ and $P^a$ all have dimensions $\mathbb{R}^{T\times 1}$.

$$P^v = Ln\left(Fl\left(Sa\left(v^F\right)\right)\right) \tag{25}$$

$$P^a = Ln\left(Fl\left(Sa\left(a^F\right)\right)\right) \tag{26}$$

TABLE I
ABLATION EXPERIMENTS ON AVE DATASET IN TWO SETTINGS

| Method | Fully Sup. $Acc.(\%)$ | Weakly Sup. $Acc.(\%)$ |
|---|---|---|
| **Fusion-stage** | | |
| $w/o$ Fusion-stage | 73.5 | 69.2 |
| $w/o$ AGVA | 74.1 | 70.4 |
| $w/o$ Bi-LSTM | 78.4 | 73.3 |
| $w/o$ Self-Att. | 77.8 | 73.1 |
| **Interaction-stage** | | |
| $w/o$ SBI | 75.0 | 71.6 |
| **Integration-stage** | | |
| $w/o$ FEE | 76.3 | 71.8 |
| CCLN(ours) | **79.9** | **75.2** |

"W/o" and "w/" indicates that the module is removed or used.

TABLE II
ABLATION EXPERIMENTS OF BRANCH MODE ON AVE DATASET

| Branch Mode | Fully Sup. $Acc.(\%)$ | Weakly Sup. $Acc.(\%)$ |
|---|---|---|
| (a) $AV$ | 78.2 | 73.7 |
| (b) $AV + V$ | **79.9** | **75.2** |
| (c) $AV + A$ | 78.6 | 74.1 |
| (d) $AV + V + A$ | 79.3 | 74.6 |

Corresponding to the modes in figure 5.

### F. Loss Design

In our CCLN framework, our weakly-supervised cross-modal contrastive learning loss (WCCL loss), introduces video-level event weak labels rather than semantics to bring related audio-visual pairs closer together. Specifically, regarding the prediction $P_i$ of one modality as the anchor, all the corresponding predictions $P_p$ with the same label within a batch are positives. $i \in I \equiv \{1 \cdots 2N\}$ is the index of any anchor within a batch, and define the index domain except anchor as $A(i) \equiv I \setminus \{i\}$. $p$ is the index of all positives with the same label within a batch, and its value domain can be written as $B(i) \equiv \{p \in A(i), y_i^w = y_p^w\}$. The other $2(N-1)$ samples within the same batch are called negatives. Our WCCL loss can be defined as,

$$\mathcal{L}_{wccl} = -\sum_{i\in I} \log \left(\frac{1}{|B(i)|} \sum_{p\in B(i)} \frac{\exp(P_i \otimes P_p/\delta) \lhd Y}{\sum_{a\in A(i)} \exp(P_i \otimes P_a/\delta) \lhd Y}\right) \tag{27}$$

where, $\delta \in \mathbb{R}^+$ is a scalar temperature parameter.

AVEL in the fully supervised setting is a segment-level multi-class classification problem. We adopt multi-class focal loss $\mathcal{L}_{mfl}$ on $P_{full}^o \in \mathbb{R}^{T\times(C+1)}$ with $Y^{full} \in \mathbb{R}^{T\times(C+1)}$ and foreground event enhancement loss (i.e. multi-class cross-entropy loss) $\mathcal{L}_{mfe}$ on $P_{full}^e \in \mathbb{R}^{T\times C}$ with $Y_{full}^e \in \mathbb{R}^{T\times C}$. In addition, we contrast the visual branch prediction $P^v$ and the fusion prediction $P^f$ are guided by label $Y_{full}^f \in \mathbb{R}^{T\times 1}$ for contrastive learning using our WCCL loss (see Table II for details). Losses

function in this setting is defined as,

$$\mathcal{L}_{mfl} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{C+1} FL\left(P^{o}_{full}, Y^{full}, \alpha, \gamma)\right) \tag{28}$$

$$\mathcal{L}_{mfe} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{C+1} Y^{e}_{full} log\left(P^{e}_{full}\right) \tag{29}$$

$$\mathcal{L}_{mcl} = \mathcal{L}_{wccl}\left(P^{v}, P^{f}, Y^{f}_{full}\right) \tag{30}$$

$$\mathcal{L}_{full} = \mathcal{L}_{mfl} + \lambda\mathcal{L}_{mfe} + \eta\mathcal{L}_{mcl} \tag{31}$$

where $FL(.)$ denotes binary focal loss.

The AVEL task can be regarded as a binary classification problem in the weakly supervised setting. We consider binary focal loss $\mathcal{L}_{bfl}$ on $P^{o}_{weak} \in \mathbb{R}^{C+1}$ with $Y^{weak} \in \mathbb{R}^{C+1}$ and foreground event enhancement loss $\mathcal{L}_{bfe}$ on $P^{e}_{weak} \in \mathbb{R}^{C}$ with $Y^{e}_{weak} \in \mathbb{R}^{C}$. Similarly, we contrast the visual branch prediction $P^{v}_{weak}$ and the fusion prediction $P^{f}_{weak}$ with label $Y^{f}_{weak}$ for contrastive learning using WCCL loss. These loss functions can be denoted as,

$$\mathcal{L}_{bfl} = -\sum_{k=1}^{C+1} FL\left(P^{o}_{weak}, Y^{weak}, \alpha, \gamma)\right) \tag{32}$$

$$\mathcal{L}_{bfe} = -\sum_{k=1}^{C+1} Y^{e}_{weak} log\left(P^{e}_{weak}\right) \tag{33}$$

$$\mathcal{L}_{bcl} = \mathcal{L}_{wccl}\left(P^{v}, P^{f}, Y^{f}_{weak}\right) \tag{34}$$

$$\mathcal{L}_{weak} = \mathcal{L}_{bfl} + \lambda\mathcal{L}_{bfe} + \eta\mathcal{L}_{bcl} \tag{35}$$

Unlike the preceding tasks, the CML model does not require a classification head. We consider binary cross-entropy loss $\mathcal{L}_{bfe}$ on $P^{o}_{cml} \in \mathbb{R}^{T\times 1}$ with $Y^{cml} \in \mathbb{R}^{T\times 1}$. The design of contrastive learning loss is the same as that of weakly supervised AVEL.

$$\mathcal{L}_{bfe} = -\sum_{k=1}^{T} Y^{cml} log\left(P^{o}_{cml}\right) \tag{36}$$

$$\mathcal{L}_{bcl} = \mathcal{L}_{wccl}\left(P^{v}, P^{f}, Y^{cml}\right) \tag{37}$$

$$\mathcal{L}_{cml} = \mathcal{L}_{bfe} + \lambda\mathcal{L}_{bcl} \tag{38}$$

## V. Experiments

### A. Experimental Descriptions

*Datasets:* The Audio-Visual Event (AVE) dataset [19], which contains 4143 samples covering 28 event categories, is used to evaluate model performance for the AVEL task. It contains audio and visual events, and the temporal boundaries of audio-visual events are manually annotated. It extensively covers real-life scenes and objects such as the church bell, frying food, train horn, toilet flush, baby cry/infant cry, etc. The AVE dataset has segment-level and video-level labels with clear temporal boundaries. The number of samples for each event category is between 60 and 188, each sample lasts 10 seconds long, and the sample is guaranteed to contain at least one audio-visual event that lasts 2 seconds long. Despite the rigorous hand-picked annotation, the AVE dataset also has problems such as content misalignment, viewpoint mutation, missing visual events, etc., but it is the most widely used large-scale audio-visual event dataset at present.

*Parameters Setup:* The video channel of the raw sample is at a frame rate of 30 fps and is 10 seconds long, cut into segments at one-second intervals. We adopt the VGG-19 network [55] pre-trained on ImageNet [56] to extract visual feature vectors. Each segment then is averaged and aggregated into one visual frame of the segment, sampling the video as a visual frame cube with dimension $\mathbb{R}^{10\times 7\times 7\times 512}$. In particular, the dimension of the visual input feature in the CML task is $\mathbb{R}^{10\times 512}$. Each segment-level audio feature is a $80-$bin log Mel filter bank, calculated by short-time Fourier transform (STFT) and fast Fourier transform (FFT). Then, the Mel filter bank features are extracted from each short-time frame, and combined with frame-level features to form a time-frequency representation. The VGG-like network [57] pre-trained on AudioSet [22] is used to extract audio feature vectors with dimension $\mathbb{R}^{10\times 128}$.

We implement our experiments on one NVIDIA Tesla V100 SXM2 under the PyTorch framework, Dropout function to regularize all the linear mappings. According to our experience, in the SBI module, we set $\tau = 0.05$ under the fully supervised setting and $\tau = 0.06$ under the weakly supervised setting. We balance the contribution of each loss by empirically selecting the optimal parameters ($\lambda = 150$ and $\eta = 25$ under the fully supervised setting, $\lambda = 100$ and $\eta = 50$ under the weakly supervised setting). Focal loss parameters $\alpha$ and $\gamma$ use default parameters. During the training stage, we set the batch size to 128 using the Adam optimizer with the default settings. There are 200 epochs throughout the training process. The initial learning rate is set at $1e-3$(under the fully supervised) and $1e-4$(under the weakly supervised). In the CML task, we set $\tau = 0.05$, $\tau = 0.05$, and the initial learning rate at $1e-4$.

*Evaluation Metric:* To make a fair comparison with other models, we follow the same evaluation metric as earlier works. Accuracy (Acc.) is often used to assess the proportion of correct predictions a model makes on the test dataset. Its metric is based on a confusion matrix and involves calculations of TP (true positive), TN (true negative), FP (false positive), and FN (false negative).

### B. Ablation Experiments

In this section, we first verify the role of each component (i.e., the Fusion-stage, Interaction-stage, and Integration-stage) of the backbone network through ablation experiments under both fully supervised and weakly supervised settings, and the experimental results are shown in Table I. For the sake of fairness and reliability of the results, we compare reproducible models, including the recent SOTA model CMBS. We use "w/o" and "w/" respectively to indicate that the component is removed or used during the experiment. Firstly, the AGVA, Bi-LSTM, and Self-Attention components of the Fusion-stage all make important contributions to the model performance, which together improve the model performance by 6.4% (fully supervised) and 6.0% (weakly supervised). In particular, among the

TABLE III
ABLATION EXPERIMENTS OF CML TASK ON AVE DATASET

| Method | A2V *Acc.*(%) | V2A *Acc.*(%) | Average *Acc.*(%) |
|---|---|---|---|
| **Fusion-stage** | | | |
| *w/o* Fusion-stage | 51.8 | 53.6 | 52.7 |
| *w/o* AGVA | 53.1 | 55.2 | 54.2 |
| *w/o* Bi-LSTM | 59.7 | 61.3 | 60.5 |
| *w/o* Self-Att. | 56.8 | 57.1 | 57.0 |
| **Interaction-stage** | | | |
| *w/o* SBI | 52.5 | 53.4 | 53.0 |
| CCLN(ours) | **63.3** | **64.4** | **63.9** |

"W/o" and "W/" indicates that the module is removed or used.

three components of the Fusion-stage, AGVA plays a relatively greater role. Secondly, the application of the Interaction-stage (SBI module) improves the model performance by about 4% (4.9%/fully supervised, 3.6%/weakly supervised). Finally, the Integration-stage (FEE module) can achieve an extra 3% gain in model performance (3.6%/fully supervised, 3.4%/weakly supervised). According to the results, the contribution of Fusion-stage is relatively higher, which may be due to the introduction of new noise in the process of cross-modal information interaction and integration. Therefore, the weights of the three stages need to find an optimal balance point to minimize the impact of noise. Although the role of the three stages of audio-visual information processing is different, we demonstrated their positive impact on model performance by ablation experiments.

Corresponding to the structural exploration of the model contrastive branch in Section III (shown in Fig. 5), we verify the rationality of the "AV+V" mode with ablation experiments. From the results in Table II, we can conclude: (1) Compared with the fusion-based mode ("AV" mode), the model performance of the "AV+V" mode are improved by 1.7% (fully supervised) and 1.5% (weakly supervised), which indicates that the two-branch structure framework can effectively reduce the noise caused by the fusion-based framework; (2) From the results of "AV+V" and "AV+A" modes, the visual modality has a greater positive effect on model performance than the audio modality; (3) The results of the "AV+V+A" mode are not optimal, combined with the limited performance improvement of the "AV+A" mode, which may be caused by the audio modality is more sensitive to noise. According to the experimental results, we finally chose the "AV+V" mode as the branch structure of CCLN.

Similarly, we conduct ablation experiments on various components of our model for the cross-modality localization task, and Table III presents the experimental results. Unlike the fully supervised and weakly supervised audio-visual event localization task, where boundaries are determined by matching predicted labels of the query modality with ground truth labels of the given modality, the CML task calculates the distance between predicted labels of the query modality and ground truth labels of the given modality to establish matching boundaries. Therefore, we removed the FEE module from the CCLN model. As shown in Table III, when the Interaction-stage design is removed from the CCLN model, the overall model performance decreases by 10.9% (from 63.9% to 53.0%). Similarly, when the Fusion-stage design is removed from the CCLN model, the overall model performance decreases by 11.2% (from 63.9% to 52.7%). The results of the ablation experiments once again confirm the significance of cross-modal fine-grained fusion design in audio-visual event localization tasks.

### C. Parameter Sensitivity Experiments

In our model, the hyperparameter $\tau$ in the SBI module is an important variable that regulates the degree of audio-visual correlation. The experimental results of $\tau$ with 0.01 intervals between 0.01 and 0.10 are shown in Table IV. We find that the model accuracy fluctuates little in fully supervised and weakly supervised settings. As the value of $\tau$ increases, it increases first and then decreases, but the optimal value is different under different settings. This means that the degree of audio-visual correlation is different for segment-level and video-level AVEL tasks. Therefore, we end up choosing $\tau = 0.05$ in the fully supervised setting and $\tau = 0.06$ in the weakly supervised setting. It is noteworthy that in the CML task, we preserve the parameters when the model achieves optimal accuracy under both "A2V" and "V2A" settings. Consequently, we solely conduct the parameter sensitivity experiment for "A2V" and selected $\tau = 0.05$.

### D. Comparison Experiments

We compare our model CCLN with the baseline AGVA [19] and the recent state-of-the-art (SOTA) models (all results are listed in Table V). Our model exceeds the accuracy results of existing models (79.9% in the fully supervised setting, 75.2% in the weakly supervised setting). In both fully supervised and weakly supervised settings, our model outperforms the baseline AGVA by 11.3% and 8.5%. Compared to the recent SOTA models, our CCLN exhibits a performance improvement of 0.4% over the SRMN [32] model in the fully supervised setting and 1.0% over the CMBS [20] model in the weakly supervised setting. Note that by designing the information processing, our model outperforms AV-Trans with more parameters (exceeds 3.1% for fully supervised and 5.0% for weakly supervised).

Fig. 8 shows the superior performance of our model in another form. It is noteworthy that all results in the figure are reproduced using the original code provided in the papers. Compared with AVGA (the baseline), our model accuracy exceeds its accuracy in 22 categories and approaches it in 4 categories. Particularly, in the "cat" category, our model accuracy outperforms the baseline by 39.4% (from 33.3% to 72.7%). Compared with CMBS (the latest SOTA model), our model accuracy exceeds its accuracy in 18 categories and approaches it in 2 categories. In the "car" category, we achieve the maximum accuracy advantage of 14.3% (from 78.9% to 93.2%). Experimental results verify that the overall performance of our model is better than other models. In addition, the accuracy of the "background" category of the three models is about 50% (49.4%/AGVA, 53.4%/CMBS, 56.4%/CCLN), which confirms the complexity of background noise and the necessity of considering it.

Table VI presents the comparative experimental results of our CCLN model with other SOTA models on the CML task of the AVE dataset. In the CML task of AVEL, only when the predicted boundaries exactly match the ground truth are they considered

TABLE IV
EFFECTS OF VARIOUS VALUES OF HYPERPARAMETER $\tau$ ON OUR MODEL ACCURACY

| $\tau$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Fully Supervised Acc.(%) | 78.4 | 78.3 | 78.7 | 79.0 | **79.9** | 79.1 | 79.3 | 78.6 | 78.8 | 78.5 |
| Weakly Supervised Acc.(%) | 72.8 | 73.2 | 73.1 | 74.1 | 74.3 | **75.2** | 74.5 | 74.4 | 73.0 | 73.9 |
| Cross-modality (A2V) Acc.(%) | 61.8 | 62.2 | 62.6 | 62.5 | **63.3** | 63.0 | 62.4 | 62.7 | 61.9 | 62.3 |

The results of the three training settings are shown.

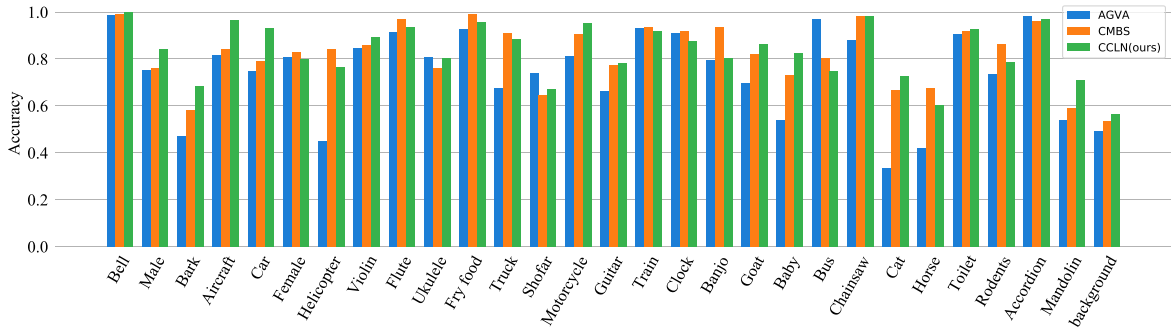



Fig. 8. Bar chart comparison of model accuracy on AVE dataset in the fully supervised setting. Each set of bar charts depicts the accuracy of AGVA (the baseline), CMBS, and our model CCLN for each audio-visual event. We show the accuracy differences between the three models across 29 event categories (including "background"). The results of the CMBS model are obtained from our reproduction of their code.

TABLE V
COMPARISON RESULTS OF OUR MODEL WITH THE SOTA MODELS FOR AUDIO-VISUAL EVENT LOCALIZATION ON THE AVE DATASET IN FULLY SUPERVISED AND WEAKLY SUPERVISED SETTINGS

| Method | Fully Sup. *Acc.*(%) | Weakly Sup. *Acc.*(%) |
|---|---|---|
| AGVA [58] | 68.6 | 66.7 |
| AVSDN [24] | 72.6 | 67.3 |
| DAM $w/o$ Matching [27] | 70.7 | - |
| DAM $w$/Self-Matching [27] | 74.2 | - |
| DAM $w$/Cross-Matching [27] | 74.5 | - |
| CMAN [29] | 73.3 | 70.4 |
| ASA [59] | 74.8 | 68.9 |
| AVIN [28] | 75.2 | 69.4 |
| CSEA [44] | 72.9 | 66.4 |
| CSPA [44] | 74.1 | 68.0 |
| CSPEA [44] | 76.5 | 70.2 |
| AV-Trans [26] | 76.8 | 70.2 |
| CMRAN $w/o$ CMRA [23] | 76.1 | 72.0 |
| CMRAN $w$/Self-Att. [23] | 76.4 | 72.5 |
| CMRAN $w$/Co-Att. [23] | 76.6 | 72.2 |
| CMRAN $w$/CMRA-F [23] | 75.6 | 71.7 |
| CMRAN $w$/CMRA [23] | 77.4 | 72.9 |
| PSP [30] | 77.8 | 73.5 |
| CMBS [20] | 79.3 | 74.2 |
| CAPB [21] | 79.3 | - |
| SRMN [32] | 79.5 | - |
| CCLN(ours) | **79.9** | **75.2** |

TABLE VI
COMPARISON RESULTS OF OUR MODEL WITH THE SOTA MODELS FOR CML ON THE AVE DATASET

| Method | A2V *Acc.*(%) | V2A *Acc.*(%) | Average *Acc.*(%) |
|---|---|---|---|
| AGVA [58] | 44.8 | 35.6 | 40.2 |
| DAM $w$/RNN [27] | 47.9 | 41.8 | 44.9 |
| DAM $w$/Avg. Pooling [27] | 46.1 | 46.0 | 46.1 |
| DAM $w$/Max Pooling [27] | 46.2 | 45.8 | 46.0 |
| DAM $w$/LSTM [27] | 48.1 | 43.5 | 45.8 |
| DAM $w$/GRU [27] | 47.4 | 45.5 | 46.5 |
| DAM $w$/Bi-LSTM [27] | 48.1 | 44.2 | 46.2 |
| DAM $w$/Self-Att. [27] | 48.5 | 47.1 | 47.8 |
| CSPA [44] | 39.3 | 33.3 | 36.3 |
| CSEA [44] | 48.5 | 50.7 | 49.6 |
| CSPEA [44] | 49.0 | 51.0 | 50.0 |
| SRMN [32] | 51.6 | 53.1 | 52.4 |
| CAPB [21] | 51.9 | 53.4 | 52.6 |
| CCLN(ours) | **63.3** | **64.4** | **63.9** |

"A2v": visual localization from audio query; "v2a": audio localization from visual query; "average": averaged accuracy of a2v and v2a.

a correct matching; otherwise, it will be deemed an incorrect matching. The percentage of correct matchings is utilized to assess the accuracy performance of the model. In Table VI, we respectively report the results for audio-to-vision (A2V) and vision-to-audio (V2A), while "Average" computes the mean accuracy for both settings. Compared to the baseline model AGVA, our CCLN model exhibits an accuracy improvement of 18.5% (A2V) and 28.8% (V2A), respectively. In comparison to the latest model CAPB, our CCLN model shows an accuracy enhancement of 11.4% (A2V) and 11.0% (V2A). This is primarily attributed to the cross-modal representation advantage brought by the dual-branch contrastive learning framework we designed.

### *E. Qualitative Analysis*

Fig. 9 shows two examples of qualitative analysis of our model. For comparison purposes, the first row of each example is a waveform image (divided into 10 segments) of the audio track with event labels, and the third row shows the ground truth (GT) frames with labels (red boxes represent the event labels). In addition, the attention heat maps of the baseline (the second row) and our model (the fourth row) are given, marking the localized event frames with blue and yellow boxes respectively. Both examples ("Goat" and "Train horn") contain background

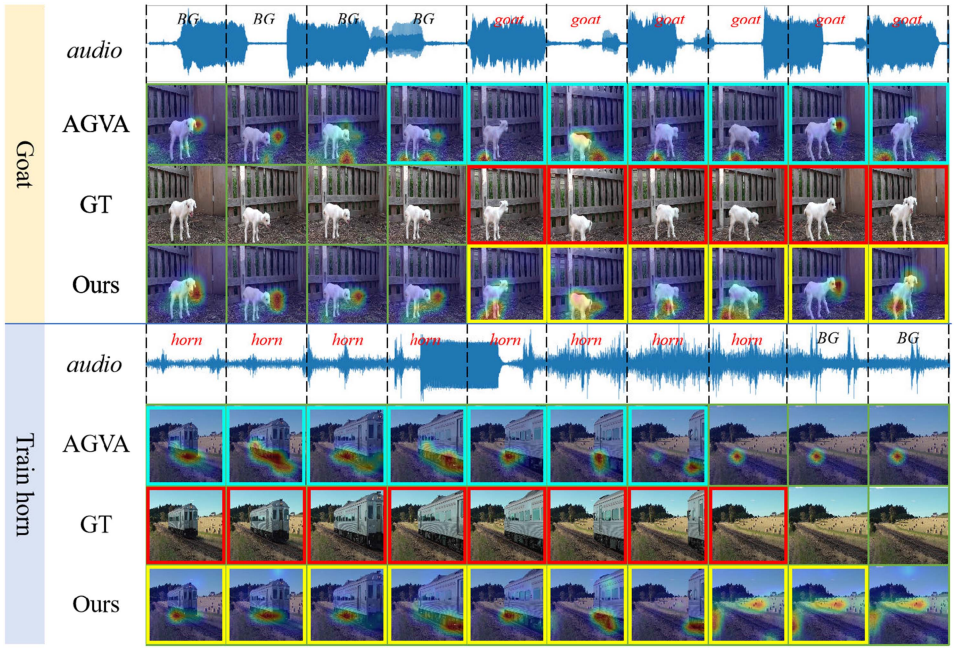
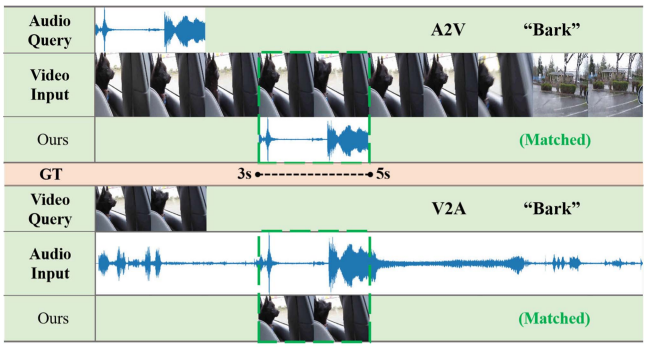
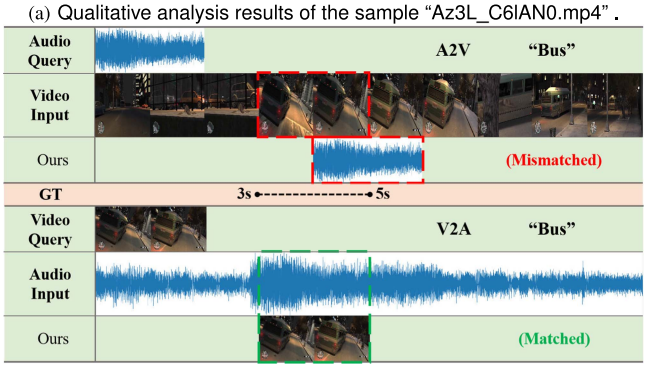



Fig. 9. Qualitative visual analysis of our model on two event examples ("Goat" and "Train horn"). Each example is divided into 10 segments, the first row of each example is a waveform image of the audio track with event labels, and the third row is visual images with the ground truth (GT) labels (the red boxes represent the event labels), and the second and fourth rows are the attention heat maps of baseline and our model (the predicted event location frames are marked with blue and yellow boxes, respectively).




(a) Qualitative analysis results of the sample "Az3L_C6lAN0.mp4".




(b) Qualitative analysis results of the sample "1zhxj3rpBcU.mp4".

Fig. 10. Qualitative results of our model on the CML task. Each example comprises two tasks: audio-to-vision (A2V) and vision-to-audio (V2A), both corresponding to the same ground truth. Green dashed boxes indicate segments where the predicted location of the query sequence matches the ground truth in the input sequence, while red dashed boxes indicate mismatches.

noise, and the second example in particular has the problem of dynamic visual multi-targets and multi-sound sources, which increases the difficulty of AVEL.

After qualitative analysis of the results in Fig. 9, we can draw the following conclusions: (1) The audio-guided visual attention (AGVA) employed alone in the baseline makes attention more sensitive to noise (the fourth frame of the "Goat" event, the eighth frame of the "Train horn" event). In other words, AGVA achieves rough cross-modal audio-visual correlation, while the noise reduction mechanism considered in our model achieves better results. (2) The attention effect of our model has a wider receptive field than the baseline (the heat area covers a larger portion of the visual target) and more accurate object localization (the heat area is closer to the visual object contour). Therefore, our model can effectively capture hidden intra and inter-modal correlations, and the refined selection of the SBI module not only expands the receptive domain but also improves the model accuracy. Of course, our model still has much room for improvement in semantic audio-visual understanding against background noise. For example, our model and baseline mislocate the seventh and ninth frames respectively, in the second example in Fig. 9. This is because, in the second example, the eighth and ninth frames are not very different in audio or visual modality, making it difficult to determine the boundaries of the audio-visual event. In addition, the richer background noise in the second example poses a greater challenge to model performance.

In the qualitative analysis of the cross-modality localization task, we perform visualizations of the audio-to-vision (A2V)

and vision-to-audio (V2A) tasks on the same test samples. In Fig. 10, green dashed boxes indicate segments where the predicted location of the query sequence matches the ground truth in the input sequence, while red dashed boxes indicate mismatches. In Fig. 10(a), despite the varying visual content, our model accurately localizes the corresponding audio query sequence. Similarly, our model successfully localizes the corresponding visual query sequence within the noisy audio input sequence. Fig. 10(b) illustrates a challenging case analysis of our model. Taking a 2-second visual segment labeled as "Bus" as the query, our model accurately matches the corresponding audio-visual event boundaries within the given audio input sequence. Conversely, when taking a 2-second audio segment labeled as "Bus" as the query, our model encounters mismatched results: the ground truth event boundaries for this sample are from the 3 rd to the 5th second, whereas our model's matched event boundaries span from the 4th to the 6th second. However, from the visualization results, it is evident that there is a high degree of similarity both the visual and audio modalities from the 3 rd to the 6th second. Therefore, distinguishing the audio-visual event boundaries between 3 to 5 seconds is a significant challenge. The continuity of visual content within the sample and the presence of background noise in the audio severely disrupt the model's ability to determine audio-visual event boundaries.

## VI. Conclusion

We propose a novel dual-branch cross-modal contrastive learning framework for AVEL. In the backbone network of the framework, we extract and validate the crucial roles of the fusion, interaction, and integration stages of audio-visual signals to explore and elucidate the mechanism of audio-visual fusion, providing a paradigm for model algorithm design. Specifically, we design a pluggable SBI module to integrate visual information, audio information, and audio-visual fusion information, and filter associated semantics through gate thresholding to further globally exploit strongly associated audio-visual events; we design an FEE module to integrate audio-visual signals at the event level, capturing event boundaries and enhancing the separation between foreground and background events. To obtain high-quality event representations in the backbone network of audio-visual fusion, we introduce a visual branch as a contrastive branch and design a weak-label-guided supervised contrastive loss function to enhance the model's representational capacity. Extensive experiments on public AVE datasets demonstrate the effectiveness of the proposed model. The results also indicate that by optimizing different stages of cross-modal information processing, the model's performance can surpass that of complex models with large-scale parameters.

## References


[1] L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe, "Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments," *Cereb. Cortex*, vol. 17, no. 5, pp. 1147–1153, 2007.

[2] M. C. Smith, *Sensory Integration: Theory and Practice*. Philadelphia, PE, USA: FA Davis, 2019.

[3] J. Jiang, A. Fares, and S.-H. Zhong, "A brain-media deep framework towards seeing imaginations inside brains," *IEEE Trans. Multimedia*, vol. 23, pp. 1454–1465, 2021.

[4] W. Zhu, X. Wang, and W. Gao, "Multimedia intelligence: When multimedia meets artificial intelligence," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1823–1835, Jul. 2020.

[5] C. Sheng, X. Zhu, H. Xu, M. Pietikäinen, and L. Liu, "Adaptive semantic-spatio-temporal graph convolutional network for lip reading," *IEEE Trans. Multimedia*, vol. 24, pp. 3545–3557, 2022.

[6] C. Sheng et al., "Importance-aware information bottleneck learning paradigm for lip reading," *IEEE Trans. Multimedia*, vol. 25, pp. 6563–6574, 2023.

[7] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *Proc. 2020 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 61–65.

[8] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vis.*, 2020, pp. 1705–1714.

[9] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 481–490.

[10] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 322–339.

[11] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 283–292.

[12] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-GCN: Correlation based graph convolutional network for audio-video emotion recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 3793–3804, 2021.

[13] M. Ren, X. Huang, W. Li, D. Song, and W. Nie, "LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 4422–4432, 2022.

[14] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Computer Vis.*, 2017, pp. 609–617.

[15] T. Afouras, Y. M. Asano, F. Fagan, A. Vedaldi, and F. Metze, "Self-supervised object detection from audio-visual correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10575–10586.

[16] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12475–12486.

[17] P. Morgado, I. Misra, and N. Vasconcelos, "Robust audio-visual instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12934–12945.

[18] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 133–142.

[19] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 247–263.

[20] Y. Xia and Z. Zhao, "Cross-modal background suppression for audio-visual event localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19989–19998.

[21] H. Wang, Z.-J. Zha, L. Li, X. Chen, and J. Luo, "Context-aware proposal–boundary network with structural consistency for audiovisual event localization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15872–15882, Nov. 2024.

[22] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 776–780.

[23] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan, "Cross-modal relation-aware networks for audio-visual event localization," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3893–3901.

[24] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang, "Dual-modality seq2seq network for audio-visual event localization," in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 2002–2006.

[25] J. Ramaswamy and S. Das, "See the sound, hear the pixels," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 2959–2968.

[26] Y.-B. Lin and Y.-C. F. Wang, "Audiovisual transformer with instance attention for audio-visual event localization," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2020, pp. 274–290.

[27] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6292–6300.

[28] J. Ramaswamy, "What makes the sound?: A dual-modality interacting network for audio-visual event localization," in *Proc. 2020 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4372–4376.
[29] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *Assoc. Advance. Artif. Intell. (AAAI)*, 2020, pp. 279–286.
[30] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang, "Positive sample propagation along the audio-visual event line," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8436–8444.
[31] C. Xue, X. Zhong, M. Cai, H. Chen, and W. Wang, "Audio-visual event localization by learning spatial and semantic co-attention," *IEEE Trans. Multimedia*, vol. 25, pp. 418–429, 2023.
[32] H. Wang, Z.-J. Zha, L. Li, X. Chen, and J. Luo, "Semantic and relation modulation for audio-visual event localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7711–7725, Jun. 2023.
[33] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1446–1453.
[34] K. W. Cheng, Y. T. Chen, and W. H. Fang, "Video anomaly detection and localization using hierarchical feature representation and gaussian process regression," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2909–2917.
[35] W. Luo, L. Wen, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. 2017 IEEE Int. Conf. Comput. Vis.*, 2017, pp. 341–349.
[36] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7834–7843.
[37] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3379–3388.
[38] J. X. Zhong et al., "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1237–1246.
[39] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6479–6488.
[40] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," 2019, *arXiv:1907.10211*.
[41] P. Wu, W. Wang, F. Chang, C. Liu, and B. Wang, "DSS-Net: Dynamic self-supervised network for video anomaly detection," *IEEE Trans. Multimedia*, vol. 26, pp. 2124–2136, 2024.
[42] C. Huang, Q. Xu, Y. Wang, Y. Wang, and Y. Zhang, "Self-supervised masking for unsupervised anomaly detection and localization," *IEEE Trans. Multimedia*, vol. 25, pp. 4426–4438, 2023.
[43] X. Min, G. Zhai, J. Zhou, M. C. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Trans. Image Process.*, vol. 29, pp. 6054–6068, 2020.
[44] C. Xue, X. Zhong, M. Cai, H. Chen, and W. Wang, "Audio-visual event localization by learning spatial and semantic co-attention," *IEEE Trans. Multimedia*, vol. 25, pp. 418–429, 2023.
[45] S. Liu et al., "Dense modality interaction network for audio-visual event localization," *IEEE Trans. Multimedia*, vol. 25, pp. 2734–2748, 2022.
[46] F. Feng, Y. Ming, N. Hu, H. Yu, and Y. Liu, "CSS-Net: A consistent segment selection network for audio-visual event localization," *IEEE Trans. Multimedia*, vol. 26, pp. 701–713, 2024.
[47] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–648.
[48] A. Zheng et al., "Adversarial-metric learning for audio-visual cross-modal matching," *IEEE Trans. Multimedia*, vol. 24, pp. 338–351, 2022.
[49] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.
[50] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K.: Springer, 2020, pp. 776–794.
[51] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn., ser. Proc. Mach. Learn. Res.*, H. D. III and A. Singh, Eds., PMLR, Jul. 2020, vol. 119, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html
[52] K. Kamnitsas et al., "Semi-supervised learning via compact latent space clustering," in *Proc. 2018 Int. Conf. Mach. Learn.*, 2018, pp. 2459–2468.
[53] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 18661–18673.
[54] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[57] S. Hershey et al., "CNN architectures for large-scale audio classification," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 131–135.
[58] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 252–268.
[59] J. Ramaswamy and S. Das, "See the sound, hear the pixels," in *Proc. 2020 IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2959–2968.