Audio-Visual Semantic Graph Network for Audio-Visual Event Localization

Liang Liu[†], Shuaiyong Li[§]*, Yongqiang Zhu[†]
[†]School of Computer Science and Technology and [§]School of Automation,
Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China
laulliam@outlook.com, lishuaiyong@cqupt.edu.cn, lnpuzyq@163.com

Abstract

Audio-visual event localization (AVEL) aims to identify both the categories and temporal boundaries of events that are both audible and visible in unconstrained videos. However, the inherent semantic gap between heterogeneous modalities often leads to audio-visual semantic inconsistency. In this paper, we propose a novel Audio-Visual Semantic Graph Network (AVSGN) to facilitate cross-modal alignment and cross-temporal interaction. Unlike previous approaches (e.g., audio-guided, visual-guided, or both), we introduce shared semantic textual labels to bridge the semantic gap between audio and visual modalities. Specifically, we present a cross-modal semantic alignment (CMSA) module to explore the complementary relationships across heterogeneous modalities (i.e., visual, audio and text), promoting the convergence of multimodal distributions into a common semantic space. Additionally, in order to capture cross-temporal dependencies sufficiently, we devise a cross-modal graph interaction (CMGI) module, which disentangles complicated interactions across modalities into three complementary subgraphs. Extensive experiments on the AVE dataset comprehensively demonstrate the superiority and effectiveness of the proposed model in both fully- and weakly-supervised AVE settings.

1. Introduction

Humans naturally integrate multiple senses (e.g., vision, hearing, and smell) to interpret their surroundings, with vision and hearing being particularly crucial [27]. To enable intelligent machines to emulate human perception, reasoning, and decision-making, audio-visual learning [36,51] has emerged as a hot area in recent decades. Audio-Visual Event Localization (AVEL) [30], a prominent area within audio-visual learning, challenges models to identify events that are both audible and visible in videos, such as a car driving or a dog barking. This task involves not only classifying event categories but also accurately determining their temporal boundaries, as shown in Figure 1. As research in AVEL continues to advance, it holds substantial potential for applications in intelligent surveillance [37], human-computer interaction [2], and multimedia retrieval [17].

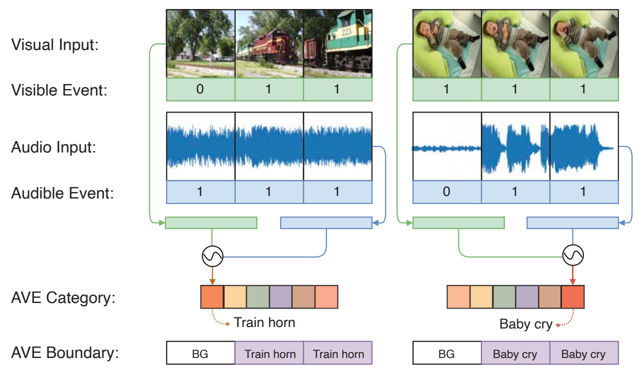

Figure 1. An example illustrating the AVEL task: a segment is marked as an "AV Event" only when an event is both visible and audible simultaneously; otherwise, it is considered background.

Since visual and audio signals often originate from the same event, the consistencies (i.e., semantic, spatial, and temporal aspects) [36] enable audio to provide valuable complementary information for AVEL tasks alongside visual data and vice versa. Since Tian et al. [30] introduced the AVEL task, extensive research have emerged to improve localization accuracy and robustness, focusing on techniques such as cross-modal attention [40, 41], multimodal adaptive fusion [4,9,16,47], and semantic consistency modeling [16,33,50]. While cross-modal attention mechanisms establish correspondences between the modalities, existing methods that rely primarily on unimodal-guided attention (e.g., AGVA [30], AGSCA [41], and CMGA [40]) often overlook the potential of visual cues to guide audio processing. As a consequence, researchers have increasingly explored the paradigm of mutual guidance [5, 7, 19, 47]. However, audio-guided and/or visual-guided attention may introduce redundant or event-irrelevant information from the other modality. Alternatively, multimodal adaptive fu-

*Corresponding author.

sion strategies aim to integrate information across modalities enhancing audio-visual correspondence through techniques like attentive fusion [4] and spatiotemporal fusion [9, 16]. Additionally, semantic consistency have been explored through dynamic interaction learning [33], consistency-guided semantic alignment [16], and consistent segment selection [7]. While these methods have shown effectiveness, they still face challenges in establishing cross-modal semantic consistency and cross-temporal dependencies due to the inherent modality heterogeneity.

In this paper, we propose a novel audio-visual semantic graph network (AVSGN) that explicitly facilitates cross-modal semantic alignment and effectively captures cross-temporal interaction between heterogeneous modalities. Inspired by the impressive zero-shot classification performance of pre-trained vision-language and audio-language models (e.g., CLIP [24], CLAP [38]), a natural idea is to introduce shared semantic labels to bridge the semantic gap between modalities. To this end, we present a cross-modal semantic alignment (CMSA) module to promote the convergence of multimodal representations into a shared semantic space. Unlike previous contrastive learning-based [34, 45], the CMSA module is entirely attention-driven, and consists of pseudo intra-modal alignment and inter-modal alignment. In addition, to suppress the interference of temporal inconsistencies, we further devise a cross-modal graph interaction (CMGI) module to capture cross-temporal semantic associations more effectively. By explicitly disentangling complicated interactions across modalities into three complementary subgraphs (i.e., audio-text, audio-visual, visual-text), we aggregate information within each subgraph to capture cross-temporal associations and achieve fine-grained fusion. Finally, we integrate these modules into the AVEL framework, achieving new state-of-the-art (SOTA) performance in both fully- and weakly-supervised AVE settings on the public AVE dataset. In summary, the contributions can be summarized as follows:

- We propose a cross-modal semantic alignment (CMSA) module that bridges the semantic gap between audio and visual modalities by introducing shared semantic labels, thereby facilitating the convergence of multimodal data into a shared semantic space.
- We devise a cross-modal graph interaction (CMGI) module that explicitly disentangles complex interactions into three complementary subgraphs to capture cross-temporal semantic associations and achieve fine-grained cross-modal temporal fusion.
- By integrating these modules into the AVEL framework, the proposed audio-visual semantic graph network (AVSGN) outperforms SOTA methods in both fully- and weakly-supervised AVE settings on the public AVE dataset.

2. Related Work

2.1. Audio-Visual Representation Learning

The inherent consistencies between audio and visual modalities provide complementary supervision that enhances audio-visual representation learning, mitigating limitations of unimodal approaches. Early methods focused on reducing modality discrepancies to improve their alignment between modalities in the context of common events or objects [3]. Recent advancements in this field have explored various strategies, including spatiotemporal correspondence, cross-modal clustering, and cross-modal contrastive learning. Synchronous-based approaches [23, 46] enhance alignment by establishing precise spatiotemporal audio-visual correspondence, thereby establishing robust modality correspondences. Clustering techniques [1, 14, 22] effectively capture complex associations and semantic consistency within audio-visual data, resulting in more representative feature extraction. Furthermore, cross-modal contrastive learning [15] has proven its effective for high-quality audio-visual representation by maximizing inter-modal similarity, showing potential when combined with pre-trained models [12]. Recently, several advanced self-supervised frameworks [28, 35, 42] have integrated the self-supervised learning and contrastive learning to effectively model the complex relationship between modalities.

2.2. Audio-Visual Event Localization

For the AVEL task, Tian et al. [30] took a pioneering step by [30] introducing an audio-guided visual attention mechanism and a dual multimodal residual network. Since then, numerous methods have been proposed to enhance the robustness and accuracy of AVEL, focusing on techniques such as cross-modal attention mechanisms [4, 11, 18, 20, 39–41, 43, 44], multimodal adaptive fusion [9, 44], and semantic consistency modeling [16, 33, 50]. Specifically, cross-modal attention mechanisms have proven its effective in extracting discriminative features from audio and visual modalities and establishing robust modality correspondence. Examples include audio-guided spatial-channel attention [41], cross-modal co-attention [20], and cross-modal gated attention [40]. Meanwhile, multimodal adaptive fusion methods integrate complementary information across modalities involving multimodal attention fusion [4] and spatiotemporal fusion encoders [9]. Semantic information plays a crucial role in further enhancing localization performance, with approaches like semantic relation modulation [33], video-level semantic consistency guidance [16], and dense modality interaction [21]. Additionally, the emergence of large pre-trained models [12] shows significant potential in AVEL [6, 19]. Recently, the audio-visual video parsing (AVVP) task [8, 29] was proposed, expanding the AVEL to detect events that are audible, visible, or both.

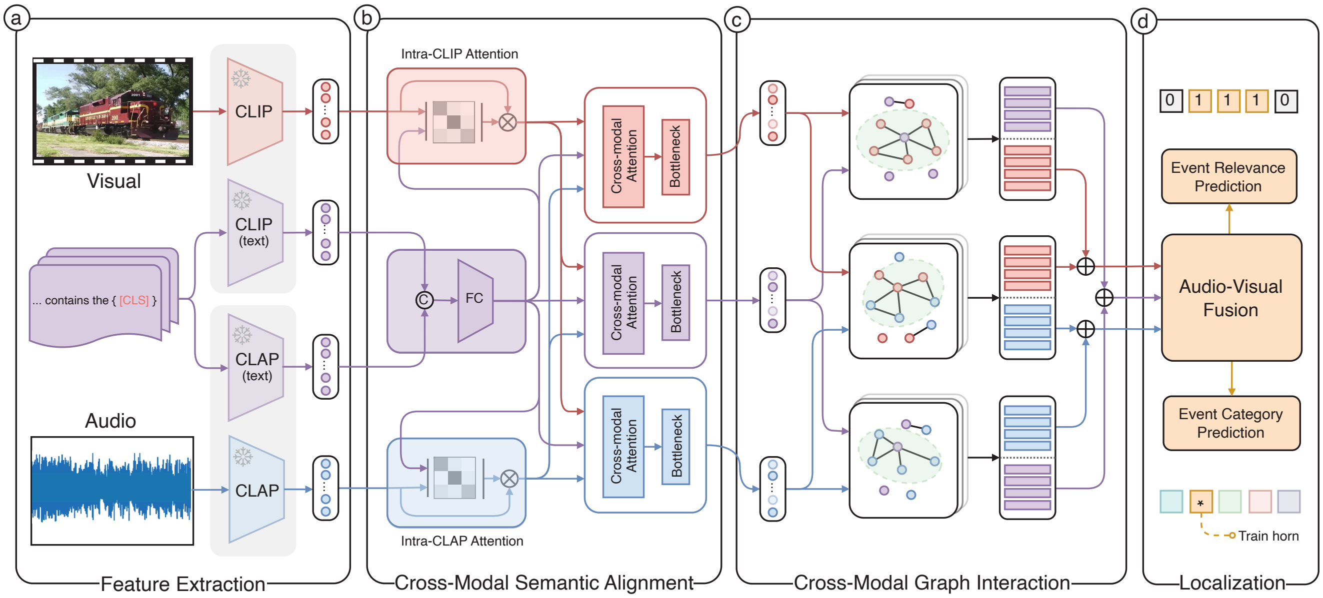

Figure 2. An overview of the proposed audio-visual semantic graph network (AVSGN). First, frozen pre-trained encoders are utilized to extract visual, audio and text embedding, respectively. Then, the cross-modal semantic alignment module aligns these multimodal information into the shared semantic space. Subsequently, three subgraphs are explicitly constructed to capture complex interactions across modalities. Finally, a localization layer predicts segment-level event relevance scores and video-level event categories.

3. Methods

3.1. Problem Formulation

The AVEL task requires identifying both the event category and its corresponding temporal boundary. Concretely, a video is divided into T sequential segments $S = \{v_t, a_t\}_{t=1}^{T}$, where the duration of each segment is 1 second, v_t and a_t represent the visual and audio features in the t-th second. The AVEL model is required to predict the event category of each segment as $y^c = \{y_t^c | y_t^c \in \{0, 1\}, \sum_{c=1}^{C+1} y_t^c = 1\}$, where $C + 1$ demotes the number of event categories (including a background category). In this paper, we primarily focus on two AVEL task settings, i.e., fully- and weakly-supervised settings. In the fully-supervised setting, the segment-level labels $y^t \in \mathbb{R}^{T\times C}$ are available during the training phase, indicating whether each segment contains an event and its corresponding event category. While in the weakly-supervised setting, only the video-level labels $y^w \in \mathbb{R}^{1\times C}$ are available during the training phase. Despite this limitation, the AVEL model is still expected to predict the event category for each segment.

3.2. Overall Pipeline

As illustrated in Figure 2, the pipeline of the proposed AVSGN model consists of feature extraction, cross-modal semantic alignment, cross-modal graph interaction, and localization. Inspired by the success of large-scale pre-trained models (e.g., CLIP [24], CLAP [38]), we employ frozen encoders to extract visual, audio and text embedding. For text encoding, the default input of CLIP follows the prompt in [49] “ `This photo contains the [CLS]`” while the default input of CLAP follows the prompt “`This sound contains the [CLS]`”. Here, the `[CLS]` represents the potential event label. Subsequently, we propose a cross-modal semantic alignment (CMSA) module to align heterogeneous context into a shared semantic space, thereby narrowing the semantic gap. Next, We devise a cross-modal graph interaction (CMGI) module, which disentangles complex inter-modal interactions into three subgraphs (i.e., audio-text, audio-visual, and visual-text) to capture cross-temporal semantic associations and achieve fine-grained cross-modal temporal fusion. Finally, we use a localization layer to predict segment-level event relevance and video-level event categories.

3.3. Cross-Modal Semantic Alignment

Although contrastive learning-based methods [13, 45] align features from different modalities, they require complex designs or constraints when dealing with three or more modalities. Unlike pervious methods, we introduce the CMSA module to explore the complementary semantic relationships among heterogeneous modalities. This module is primarily attention-based and consists of pseudo intra-modal alignment and inter-modal alignment.

Pseudo Intra-modal Alignment: For multimodal AVEL task, information from different modalities (e.g., visual and text) may contain complementary event-related information. Although visual and text data belong to different modalities, CLIP [24] (CLAP [38]) is pretrained on a large-scale dataset of paired image-text (audio-text) samples, which allows it to learn implicit associations between these modalities. Due to the differences in text encoding between CLIP and CLAP, we concatenate the text embedding generated by both encoders and then project them into a shared space to ensure consistency in sequence length between visual, audio, and text embedding. The above operations can be expressed as: $\mathcal{T} = \Phi_t([\mathcal{T}_{CLIP}; \mathcal{T}_{CLAP}])$, where $\mathcal{T} \in \mathbb{R}^{T\times d}$ is global label text embedding, $\Phi_t(\cdot) \in \mathbb{R}^{T\times d}$ is a linear layer, and $\mathcal{T}_{CLIP} \in \mathbb{R}^{C\times d}$ and $\mathcal{T}_{CLAP} \in \mathbb{R}^{C\times d}$ are text embedding encoded by CLIP and CLAP text encoder. In order to further enhance potential event associations across modalities, we introduce intra-CLIP attention and intra-CLAP attention mechanisms to pay more attention to event-relevant information. The standard attention mechanism [32] can be formulated as:

$$\text{Attn}(Q, K, V) = \text{softmax}(\frac{(QW_q)(KW_k)^T}{\sqrt{d}})(VW_q) \quad (1)$$

where W_q, W_k and W_v are trainable parameters, with $\sqrt{d}$ serving as the scaling factor. Given video embedding $\mathcal{V} \in \mathbb{R}^{T\times d}$ and label embedding $\mathcal{T} \in \mathbb{R}^{T\times d}$, the intra-CLIP visual attention $\mathcal{V}_{attn} \in \mathbb{R}^{T\times d}$ is defined as:

$$\mathcal{V}_{attn} = \text{Attn}(\mathcal{T}, \mathcal{V}, \mathcal{V}) \quad (2)$$

where the label embeddings serve as the query, while the visual embeddings act as the key and value. Likewise, the intra-CLAP audio attention $\mathcal{A}_{attn} \in \mathbb{R}^{T\times d}$ is defined as:

$$\mathcal{A}_{attn} = \text{Attn}(\mathcal{T}, \mathcal{A}, \mathcal{A}) \quad (3)$$

where we designate the label embedding as query (Q) and audio embedding $\mathcal{A} \in \mathbb{R}^{T\times d}$ as key (K) and value (V). Furthermore, we extend both the intra-CLIP visual attention and intra-CLAP audio attention to multi-head attention, allowing the model to capture diverse semantic aspects across multiple subspaces.

Inter-modal Alignment: To bring the target modality closer to the other two modalities, we leverage the attention mechanism [48] to pull heterogeneous context into a shared semantic space. As shown in Figure 3, taking visual alignment as example, we first calculate the attention weights between different elements across modalities, and then concatenate corresponding outputs, and pass them to a bottleneck structure to generate aligned visual features. Consequently, the aligned visual attention $\mathcal{F}_{v,attn} \in \mathbb{R}^{T\times d}$ can be formulated as:

$$\mathcal{F}_{v,attn} = \Phi_v([\text{Attn}(\mathcal{A}, \mathcal{V}, \mathcal{V}); \text{Attn}(\mathcal{T}, \mathcal{V}, \mathcal{V})]) \quad (4)$$

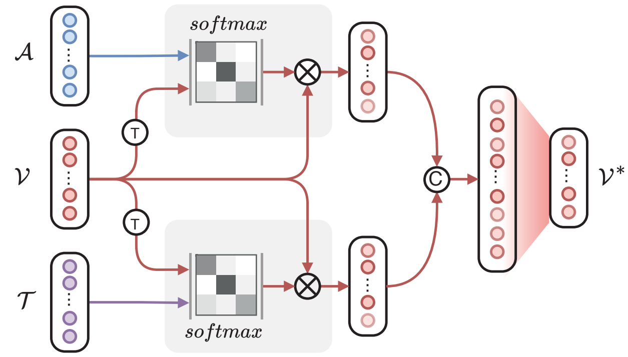

Figure 3. An illustration of cross-modal alignment, taking visual alignment as example.

where $\Phi_v(\cdot) \in \mathbb{R}^{T\times d_h}$ is a fully-connected layer with ReLU activation, $\mathcal{V}_{attn}$ and $\mathcal{V}$, as well as $\mathcal{A}_{attn}$ and $\mathcal{A}$ are equivalent in the current context. The aligned visual representations aggregate event-relevant information from both audio and text modalities, effectively bridging the gap across modalities within shared semantic space. Likewise, we can obtain the aligned audio attention $\mathcal{F}_{a,attn} \in \mathbb{R}^{T\times d_h}$ and text attention $\mathcal{F}_{t,attn} \in \mathbb{R}^{T\times d_h}$ by:

$$\mathcal{F}_{a,attn} = \Phi_a([\text{Attn}(\mathcal{V}, \mathcal{A}, \mathcal{A}); \text{Attn}(\mathcal{T}, \mathcal{A}, \mathcal{A})]) \quad (5)$$

$$\mathcal{F}_{t,attn} = \Phi_t([\text{Attn}(\mathcal{A}, \mathcal{T}, \mathcal{T}); \text{Attn}(\mathcal{V}, \mathcal{T}, \mathcal{T})]) \quad (6)$$

where $\Phi_a(\cdot) \in \mathbb{R}^{T\times d_h}$ and $\Phi_t(\cdot) \in \mathbb{R}^{T\times d_h}$ are nonlinear layers with ReLU activation function. Through this process, we achieve semantic alignment across heterogeneous modalities, thereby effectively bridging the semantic gap among heterogeneous data within the feature space.

3.4. Cross-Modal Graph Interaction

To effectively capture the dynamic dependencies between multiple modalities, we devise a cross-modal graph interaction (CMGI) module. We explicitly disentangle complex interactions across modalities into three complementary subgraphs, i.e., audio-text $G_{at} = (\mathcal{V}_{at}, \mathcal{E}_{at})$, audio-visual $G_{av} = (\mathcal{V}_{av}, \mathcal{E}_{av})$, and visual-text $G_{vt} = (\mathcal{V}_{vt}, \mathcal{E}_{vt})$. Here, $\mathcal{V}$ represents set of graph nodes, and $\mathcal{E}$ represents the set of graph edges.

Taking the subgraph G_{av} as an example, both audio-text subgraph and visual-text subgraph are implemented analogously. In the subgraph G_{av}, visual/audio features of each moment are regarded as a node. Since adjacency relationship is crucial for graph, we apply the attention mechanism to make the relationship between each pair of modal nodes, which not only rely on fixed structural information but dynamically adjust based on semantic similarity. Given visual features $\mathcal{F}_{v,attn} \in \mathbb{R}^{T\times d_h}$ and audio features $\mathcal{F}_{a,attn} \in \mathbb{R}^{T\times d_h}$ in semantic space, we use attention mechanism to update the weights between nodes. Firstly, we concatenate the audio and visual features along the node dimension to form the initial node features $X_{av} = [\mathcal{F}_{a,attn}; \mathcal{F}_{v,attn}] \in \mathbb{R}^{2T\times d_h}$. Thus, the adjacency matrix, or attention weights, $e_{av} \in \mathbb{R}^{2T\times 2T}$ can be formulated as:

$$e_{av} = \text{softmax}(\frac{(X_{av}W_{av}^q)(X_{av}W_{av}^k)^T}{\sqrt{d_h}}) \tag{7}$$

where $W_{av}^q \in \mathbb{R}^{d_h\times d_h}$ and $W_{av}^q \in \mathbb{R}^{d_h\times d_h}$ are learnable query matrix, and key matrix, respectively. Similarly, the attention weights of subgraph G_{vt} and G_{at} can be represented as $e_{vt} \in \mathbb{R}^{2T\times d_h}$ and $e_{at} \in \mathbb{R}^{2T\times d_h}$, respectively.

Subsequently, we feed the attention weights e_{av} and node features X_{av} into the multilayer graph convolutional network (GCN) to further propagate information and aggregate the node representations. Specifically, each GCN layer uses the connection relationship in the adjacency matrix to pass the local information of the node to the neighbor nodes. For each node i in subgraph G_{av}, the node representation at layer $l+1$ can be defined as:

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \tilde{A}_{ij}X_j^{(l)}W^{(l)}\right) \tag{8}$$

where $h_i^{(l)}$ is the feature vector of node i at l layer, $\mathcal{N}(i)$ is a collection of neighbors for node i, $\tilde{A}_{ij}$ is the (i,j)-th entry of the adjacency matrix $\tilde{A} \in \{e_{av}, e_{vt}, e_{at}\}$, $W^{(l)}$ is the learnable weight matrix for l layer, σ is a nonlinear activation function, e.g., ReLU.

After multiple iterations of GCN layers, we obtain the modality-specific representations from the three subgraphs: $\mathcal{A}_{at} \in \mathbb{R}^{T\times d_h}$ and $\mathcal{V}_{av} \in \mathbb{R}^{T\times d_h}$, audio and text representation on G_{at}: $\mathcal{A}_{at} \in \mathbb{R}^{T\times d_h}$ and $\mathcal{T}_{at} \in \mathbb{R}^{T\times d_h}$, and visual and text representation on G_{vt}: $\mathcal{V}_{at} \in \mathbb{R}^{T\times d_h}$ and $\mathcal{T}_{at} \in \mathbb{R}^{T\times d_h}$. We then sum the features of the corresponding modalities to obtain the final representations: audio representation $\mathcal{A}^* = \mathcal{A}_{av} \oplus \mathcal{A}_{at}$, visual representation $\mathcal{V}^* = \mathcal{V}_{av} \oplus \mathcal{V}_{vt}$, and text representation $\mathcal{T}^* = \mathcal{T}_{at} \oplus \mathcal{T}_{vt}$. By structuring interactions through dedicated three complementary subgraphs, the CMGI module effectively alleviates temporal inconsistency and captures cross-temporal event cues across modalities.

3.5. Classification and Objective Function

The above modules generate discriminative embedding for both the audio and visual modalities, as well as text label embedding, to be used by the subsequent classifier. Inspired by [40], we introduce a gated mechanism to integrate complementary information. Concretely, we add two gates for audio and visual: $g_a = \sigma(W_g^a\mathcal{A}^*)$ and $g_v = \sigma(W_g^v\mathcal{V}^*)$, where $W_g^a \in \mathbb{R}^{d_h\times 1}$ and $W_g^v \in \mathbb{R}^{d_h\times 1}$ are learnable parameters. Thus, the joint audio-visual representation $\mathcal{F}^o \in \mathbb{R}^{T\times d_h}$ is depicted as:

$$\mathcal{F}^o = (\mathcal{V}^* \odot g_a) \oplus (\mathcal{A}^* \odot g_v) \tag{9}$$

Finally, the classification details for fully- and weakly-supervised settings will be introduced in detail below.

Fully-supervised AVEL task: In line with mainstream AVEL methods (e.g., [40, 41]), we divide the fully-supervised task into two subtasks, i.e., video-level category prediction and segment-level relevance prediction. Specifically, the video-level category score s^c and segment-level relevance score s^t are defined as follows:

$$s^c = \text{softmax}(\text{max-pooling}(\mathcal{F}^o)\cdot W_c) \tag{10}$$

$$s^t = \sigma(\mathcal{F}^o \cdot W_t) \tag{11}$$

where $W_c \in \mathbb{R}^{d\times C}$ and $W_t \in \mathbb{R}^{d_h\times 1}$ are the learnable parameters, σ is the sigmoid activation function. Thus, the event-relevant loss $\mathcal{L}_r$ and event category loss $\mathcal{L}_c$ can be formulated as:

$$\mathcal{L}_c = CE(y^c, s^c) \tag{12}$$

$$\mathcal{L}_r = BCE(y^r, \mathbb{I}(s^t \geq \tau)) \tag{13}$$

where $CE(,)$ represents cross-entropy, $BCE(,)$ denotes binary cross-entropy, $\mathbb{I}(\cdot)$ is an indicator function, y^c and y^r are the ground truth, τ is a threshold to distinguish whether a segment is an event-related segment or background. Additionally, to ensure that the gates g_a, g_v capture intrinsic audio-visual patterns, we use ground truth category label y^c to optimize network and introduce the gate loss $\mathcal{L}_g$:

$$\mathcal{L}_g = CE(y^c, (g_a \odot g_v)\cdot W_g^{av}) \tag{14}$$

where $W_g^{av} \in \mathbb{R}^{d_h\times C}$ represents the fully-connected layer.

Furthermore, we introduce semantic consistency loss to ensure audio-visual representations aligned within the semantic space. The semantic consistency loss at the t-th segment of a video is defined as:

$$\mathcal{L}_s^t = y^t\|v_t - a_t\| + (1-y^t)\cdot\max(0, m - \|v_t - a_t\|) \tag{15}$$

where $v_t \in \mathcal{V}^*$, $a_t \in \mathcal{A}^*$, m is a margin controlling the distance of extraneous segments, and y^t denotes the event-relevant label for the t-th segment. Therefore, the final loss function of fully-supervised localization can be defined as:

$$\mathcal{L}_{fully} = \mathcal{L}_r + \mathcal{L}_c + \mathcal{L}_g + \lambda\mathcal{L}_s \tag{16}$$

where $\mathcal{L}_s = \sum \mathcal{L}_s^t$ and λ denotes a hyperparameter.

Weakly-supervised AVE task: We formulate weakly-supervised AVEL as a multiple instance learning (MIL) task. Since only video-level event labels are available during the training phase, we cannot utilize the ground truth segment-level label to optimize the model. In this case, the event-relevant score s^t and event category score s^c are all first expanded to the dimension of $\mathbb{R}^{T\times C}$. Then, we fuse s^t and s^c to yield the joint score $s^{ct} \in \mathbb{R}^{T\times C}$. Thus, the MIL loss can be written as:

$$\mathcal{L}_{MIL} = CE(y^c, MIL(s^{ct})) \tag{17}$$

where $MIL(\cdot) \in \mathbb{R}^C$ denotes the multiple instance learning pooling. Similar to the fully-supervised setting, we introduce the gate loss to capture underlying audio-visual patterns. Finally, the loss function of weakly-supervised localization can be expressed as:

$$\mathcal{L}_{weak} = \mathcal{L}_{MIL} + \mathcal{L}_g \tag{18}$$

4. Experiments

4.1. Experimental Setup

Dataset: The AVE dataset [30] derived from AudioSet [10], contains 4,143 videos and 28 categories, such as dog barking, helicopter, and acoustic guitar. Each video sample is divided into 10 one-second segments. In addition, each video in the dataset contains segment-level AVE labels and video-level event category labels.

Evaluation Metrics: For the AVEL task, the event category label of each segment is required to be predicted in both fully- and weakly-supervised settings. Following [40], we adopt the overall classification accuracy of the predicted segment-level event category as the evaluation metric.

Implementation Details: We utilize the Swin-V2-L-based CLIP [24] and HTS-AT-based CLAP [38] to obtain 512D visual, audio, and label text embedding, respectively. We employ Adam optimizer with a batch size of 64 to train the model, where τ defaults to 0.6 and λ defaults to 0.01. We initialize the learning rate to 3×10^{-4} and is halved at epoch 10, 20, and 30. Our model is trained with 200 epoch and evaluated on an NVIDIA GeForce RTX 2080 Ti GPU and implemented in Pytorch.

4.2. Comparisons with SOTA Methods

Fully-supervised AVEL: As is shown in Table 1, our method achieves a new SOTA performance (83.1%) in comparison to the current methods. Notably, our method significantly surpasses baseline models with feature extractors (VGG-19, VGG-like), e.g., CMBS [40] (relatively +3.8%), CSS-Net [7] (relatively +2.6%). These results not only underscore the effectiveness of our model but demonstrate the powerful backbone (i.e., feature extractor) is crucial to extract compact and representative features indirectly. In addition, we also benchmark our approach against methods that similarly utilize frozen large pre-trained models, e.g., LAVISH [19], DG-SCT [6]. According to the experimental results, the proposed method outperforms the previous LAVISH [19] by 4.5% and suppresses the suboptimal DG-SCT [6] by 0.9%, respectively. Therefore, the above results further confirm that both our model design and the utilization of a strong feature extractor contribute significantly to setting a new benchmark in AVEL performance.

Table 1. Performance (%) comparison with SOTA methods in the fully- and weakly-supervised settings on the AVE dataset.

Models	Encoder	F. Sup.	W. Sup.
AVEL [30]	VGG-19, VGG-like	72.7	66.7
AVSDN [18]	VGG-19, VGG-like	72.6	67.3
CMAN [43]	VGG-19, VGG-like	73.3	70.4
DAM [39]	VGG-19, VGG-like	74.5	–
AVRB [26]	VGG-19, VGG-like	74.8	68.9
AVIN [25]	VGG-19, VGG-like	75.2	69.4
CSA [44]	VGG-19, VGG-like	76.5	70.2
AVT [20]	VGG-19, VGG-like	76.8	70.2
CMRAN [41]	VGG-19, VGG-like	77.4	72.9
PSP [50]	VGG-19, VGG-like	77.8	73.5
CMBS [40]	VGG-19, VGG-like	79.3	74.2
DMIN [21]	VGG-19, VGG-like	79.6	74.3
VSCG [16]	VGG-19, VGG-like	79.7	74.8
CSS-Net [7]	VGG-19, VGG-like	80.5	76.8
LAVISH$^{\triangle}$ [19]	Swin-V2-L, HTS-AT	78.6	–
DG-SCT$^{\triangle}$ [6]	Swin-V2-L, HTS-AT	82.2	–
AVSGN (Ours)	Swin-V2-L, HTS-AT	**83.1**	**78.8**

[1] Note: $\triangle$ indicates results obtained from [6].

Weakly-supervised AVEL: Table 1 also shows how our method compares to current methods on the weakly-supervised setting. Compared to current methods, our method achieves a new SOTA performance (78.8%), further demonstrating the effectiveness of our model in both settings. Notably, our method significantly outperforms the baseline models with feature extractors (VGG-19, VGG-like), including CMBS [40] (relatively +4.6%) and CSS-Net [7] (relatively +2.0%). By the way, LAVISH [19] and DG-SCT [6] did not report results on the weakly-supervised setting. In summary, despite the absence of segment-level labels, our method achieves remarkable performance, even outperforming some early fully-supervised methods, such as CMRAN [41] and AVT [20].

4.3. Ablation Studies

1) *Does the Cross-modal Semantic Alignment help?* To estimate the effectiveness of CMSA, we selectively remove components from CMSA module. As shown in Table 2, Ω_{CLIP} and Ω_{CLAP} represent the intra-CLIP and intra-CLAP alignment, and $\mho_{\text{V}}$, $\mho_{\text{A}}$, and $\mho_{\text{T}}$ denote cross-modal alignment for visual, audio, and label text, respectively. According to the experimental results, when Ω_{CLIP} and Ω_{CLAP} are removed, we observe a slight decrease in both fully- and weakly-supervised settings. This slight decrease highlights that pseudo intra-modal alignment captures the potential correlation between visual (audio) and label em-

Table 2. Ablation studies on the effectiveness of pseudo intra-modal and inter-modal alignment in the CMSA module.

Ω_{CLIP}	Ω_{CLAP}	$\mho_{\mathrm{V}}$	$\mho_{\mathrm{A}}$	$\mho_{\mathrm{T}}$	**F. Sup.**	**W. Sup.**
✓	✓	✓	✓	✓	83.1	78.8
×	×	✓	✓	✓	82.8	78.5
✓	✓	×	✓	✓	80.4	76.0
✓	✓	✓	×	✓	81.8	77.2
✓	✓	✓	✓	×	82.3	77.8
✓	✓	×	×	×	79.5	75.4
×	×	×	×	×	79.0	74.5

[1] Note: Ω indicates pseudo intra-modal alignment, while $\mho$ denotes the cross-modal alignment.

Table 3. Ablation studies on the effectiveness of three subgraph structures in the CMGI module.

A↔V	V↔T	A↔T	**F. Sup.**	**W. Sup.**
✓	✓	✓	83.1	78.8
×	✓	✓	80.3	75.9
✓	×	✓	81.7	76.6
✓	✓	×	82.5	78.0
×	×	×	79.6	74.8

Table 4. Ablation studies on the effectiveness of loss functions on fully-supervised setting.

$\mathcal{L}_c$	$\mathcal{L}_r$	$\mathcal{L}_g$	$\mathcal{L}_s$	**F. Sup.**
✓	✓	✓	✓	83.1
×	×	✓	✓	79.5
✓	✓	×	✓	81.7
✓	✓	✓	×	82.0
✓	✓	×	×	80.2

bedding, contributing positively to the model. In addition, when cross-modal visual alignment $\mho_{\mathrm{V}}$ is removed, the performance significantly drops (nearly -3%) in both fully- and weakly-supervised settings, which indicates that cross-modal visual alignment plays a critical role in capturing event semantics effectively. Similarly, removing $\mho_{\mathrm{A}}$ (alignment of audio features) results in a considerable performance decrease, though less pronounced than removing $\mho_{\mathrm{V}}$, which indicates that cross-modal audio alignment is also important. Compared to cross-modal visual and audio alignment, $\mho_{\mathrm{T}}$ also positively impacts performance but has less influence, indicating a supporting role in the cross-modal alignment process.

2) *Does the Cross-modal Graph Interaction help?* Analogously, we validate the effectiveness of different components of the proposed CMGI module. In the CMCI module, we explicitly disentangle complicated interactions across modalities into three complementary subgraphs, i.e., audio-visual (A↔V), visual-text (V↔T), and audio-text (A↔T).

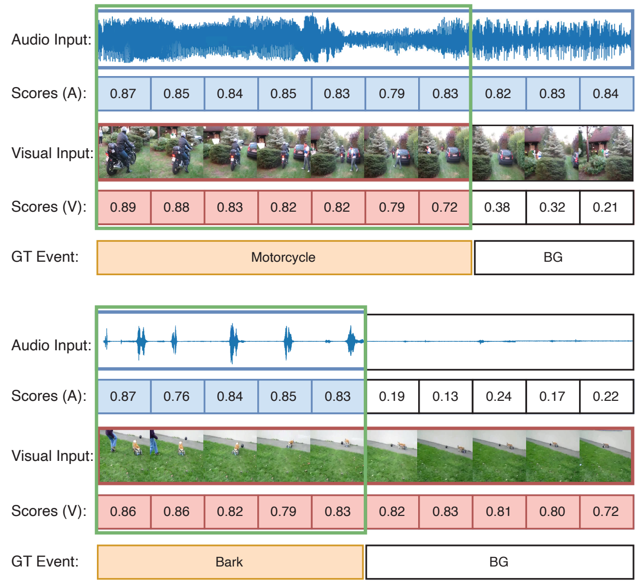

Figure 4. Qualitative results of our model on the *motorcycle* event (top) and the *dog bark* event (bottom).

Table 3 demonstrates the effectiveness of each cross-modal interaction component within our CMGI module. Concretely, removing A↔V component results in the most significant performance drop, which indicates that direct interaction between audio and visual features is crucial, as it potentially captures the primary cues for audio-visual synchronization and co-occurring events. When the V↔T interaction is removed, performance also drops, which indicates that interaction between visual and textual information is also vital, as it helps contextualize visual scenes with event-relevant text features. Although A↔T interaction is less effective than the others, it still contributes to enhanced performance to some extent, likely by reinforcing the semantic relationship between audio events and label embedding. Overall, the CMGI module plays a positive role in the model and shows potential in alleviating temporal inconsistency between modalities through capturing cross-temporal event cues.

3) *Does the different loss functions help?* To explore the effectiveness of loss functions (i.e., event category loss $\mathcal{L}_c$, event-relevant loss $\mathcal{L}_r$, audio-visual gate loss $\mathcal{L}_g$, and semantic consistency loss $\mathcal{L}_s$), we conduct the ablation study, as shown in Table 4. The video-level classification loss $\mathcal{L}_c$ and segment-level event relevance loss $\mathcal{L}_r$, commonly used in mainstream methods [40, 41], as removing them leads to a significant drop in performance. This observation underscores their critical contribution to capture segment-level and video-level event cues. In addition, when we remove the gating loss $\mathcal{L}_g$, the performance also drops, indicating that the gating mechanism helps refine multi-modal inter-

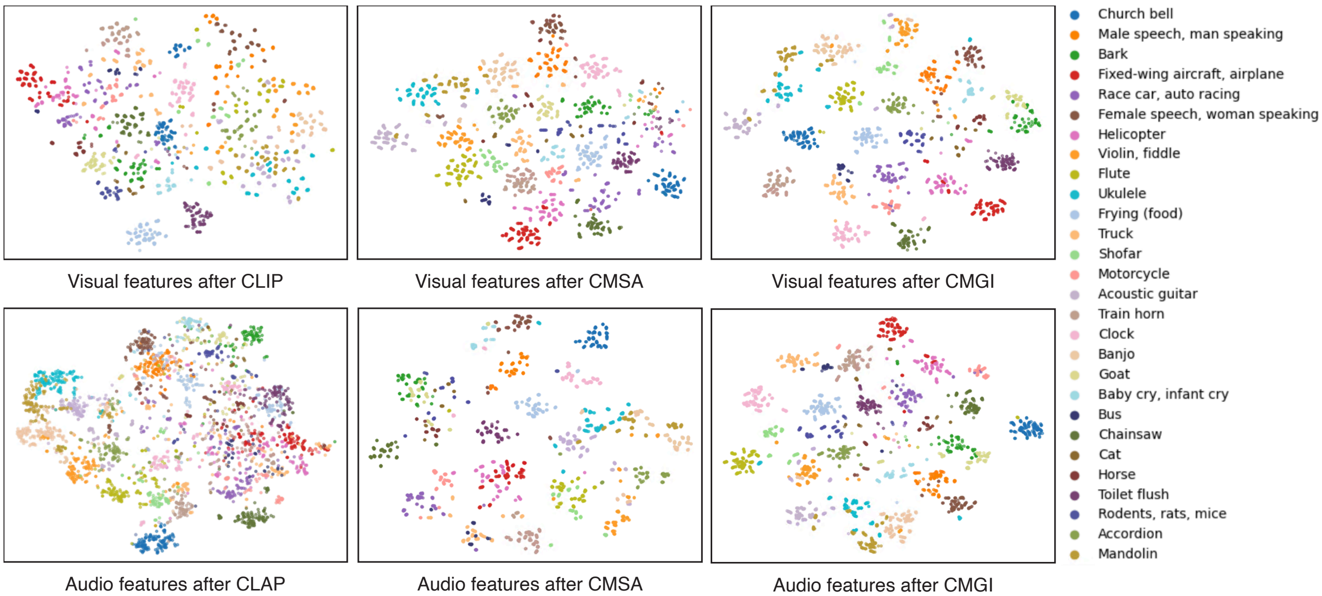

Figure 5. Illustration of feature distributions using t-SNE [31] in fully-supervised learning. The first row shows the distribution of audio features after different components, while the second row shows the distribution of visual features.

actions and strengthens the overall prediction accuracy by adaptively filtering out event-irrelevant noise. Furthermore, without the semantic consistency loss $\mathcal{L}_s$, the performance experiences a notable decrease, indicating the importance of ensuring the audio-visual consistency at the segment level.

4.4. Qualitative Analysis

As illustrated in Figure 4, we present qualitative results for two examples. Each audio and visual segment receives a score calculated based on individual relevance scores s_r and gate values g_a or g_v, respectively. Specifically, in the first example ("Motorcycle"), continuous audio alone doesn't identify the event boundary, but visual changes help pinpoint it, allowing the model to distinguish the "Motorcycle" event and background. In the second example ("Bark"), the consistent visual input provides limited information, so the model relies on the audio modality to detect the event boundary. Furthermore, we also visualize the feature distribution after processing by the CLIP [24] (CLAP [38]), the proposed CMSA module, and the proposed CMGI module. Initially, features obtained from the pre-trained models (CLIP and CLAP) appear somewhat dispersed and lack clear discrimination between different classes. However, after applying the CMSA module, the both audio and visual features become discriminative, reflecting the CMSA module's ability to capture semantic relationships across modalities. In addition, the audio/visual features of different categories are more compact after applying CMGI. Notably, semantically related categories, such as musical instruments (e.g., Flute and Ukulele) and human speech (e.g., Male speech and Female speech), are mapped closer together in the feature space, further validating the effectiveness of our method in capturing semantic associations.

5. Conclusion

In this paper, we introduce a novel audio-visual semantic graph network (AVSGN) to alleviate the audio-visual inconsistency in AVEL task through cross-modal semantic alignment and cross-temporal graph interaction between heterogeneous modalities. The cross-modal semantic alignment (CMSA) module, consisting of pseudo intra-modal alignment and inter-modal alignment, enhances alignment across different modalities and contributes to promote the convergence of multimodal distributions into a shared semantic space. In addition, the complicated associations across modalities are explicitly disentangled into three complementary subgraphs (i.e., audio-text, audio-visual, and visual-text). To capture cross-temporal semantic interactions sufficiently, we devise a cross-modal graph interaction (CMGI) module to suppress the interference of temporal inconsistencies. The experimental results demonstrate the superiority of the proposed AVSGN model over current SOTA approaches in both fully- and weakly-supervised settings. Also, the ablation studies demonstrate that our approach produces more compact and discriminative representations, which further validate the effectiveness of our method. Furthermore, the framework can be extended to other related multimodal tasks (i.e., audio-visual video parsing), where the advantages of a graph-based model could more effectively handle the multi-modal, multi-instance task.

Acknowledgments

This work was supported by the Chongqing Natural Science Foundation Innovation and Development Joint Foundation (CSTB2024NSCQ-LZX0035), Science and Technology Research Project of Chongqing Education Commission (KJZD-M202300605), New Chongqing Youth Innovation Talent Plan Project (CSTB2024NSCQ-QCXMX0053), Special general project for Chongqing's technological innovation and application development (CSTB2024TIAD-KPX0073, CSTB2024TIAD-KPX0101, CSTB2024TIAD-KPX0027), Yongjiang Program for Young Talents (RC20230107), Chongqing Research Institution Performance Incentive Guidance Special Project (CSTB2023JXJL-YFX0013), Chongqing University of Posts and Telecommunications Ph.D. Innovative Talent Project (BYJS202409).

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020. 2

[2] Josh Andres, Nathan Semertzidis, Zhuying Li, Yan Wang, and Florian Floyd Mueller. Integrated exertion—understanding the design of human–computer integration in an exertion context. *ACM Trans. Comput.-Hum. Interact.*, 29(6), 2023. 1

[3] Relja Arandjelovic and Andrew Zisserman. Look, Listen and Learn. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 609–617, 2017. 2

[4] Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont. Multimodal attentive fusion network for audio-visual event recognition. *Inf. Fusion*, 85(C):52–59, 2022. 1, 2

[5] Tingxiu Chen, Jianqin Yin, and Jin Tang. Past and future motion guided network for audio visual event localization. *arXiv e-prints*, 2022. 1

[6] Haoyi Duan, Yan Xia, Mingze Zhou, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: adapting large pre-trained models for audio-visual downstream tasks. In *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, 2024. 2, 6

[7] Fan Feng, Yue Ming, Nannan Hu, Hui Yu, and Yuanan Liu. CSS-Net: A consistent segment selection network for audio-visual event localization. *IEEE Transactions on Multimedia*, 26:701–713, 2024. 1, 2, 6

[8] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 18827–18836, 2023. 2

[9] Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. Learning event-specific localization preferences for audio-visual event localization. In *Proc. 31st ACM Int. Conf Multimedia*, pages 3446–3454, 2023. 1, 2

[10] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. AudioSet: An ontology and human-labeled dataset for audio events. In *Proc. Int. Conf. Acoust. Speech, Signal Processing.*, pages 776–780, 2017. 6

[11] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 22942–22951, 2023. 2

[12] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *Proc. Int. Conf. Acoust. Speech, Signal Processing.*, pages 976–980, 2022. 2

[13] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 14867–14878, 2023. 3

[14] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 9240–9249, 2019. 2

[15] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, haoqi fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. Mavil: Masked audio-video learners. In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 20371–20393, 2023. 2

[16] Yuanyuan Jiang, Jianqin Yin, and Yonghao Dang. Leveraging the video-level semantic consistency of event for audio-visual event localization. *IEEE Trans. Multimedia*, 26:4617–4627, 2023. 1, 2, 6

[17] Cong Jin, Tian Zhang, Shouxun Liu, Yun Tie, Xin Lv, Jianguang Li, Wencai Yan, Ming Yan, Qian Xu, Yicong Guan, and Zhenggougou Yang. Cross-modal deep learning applications: Audio-visual retrieval. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 301–313, 2021. 1

[18] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *Proc. Int. Conf. Acoust. Speech, Signal Processing.*, pages 2002–2006, 2019. 2, 6

[19] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 2299–2309, 2023. 1, 2, 6

[20] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proc. Asian Conf. Comput. Vis.*, pages 274–290, 2021. 2, 6

[21] Shuo Liu, Weize Quan, Chaoqun Wang, Yuan Liu, Bin Liu, and Dong-Ming Yan. Dense modality interaction network for audio-visual event localization. *IEEE Trans. Multimedia*, 25:2734–2748, 2023. 2, 6

[22] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 10565–10574, 2023. 2

[23] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement.

In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 12470–12481, 2021. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 6, 8

[25] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *Proc. Int. Conf. Acoust. Speech, Signal Processing.*, pages 4372–4376, 2020. 6

[26] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 2959–2968, 2020. 6

[27] Patricia A. Reuter-Lorenz, Kathleen Baynes, George R. Mangun, and Elizabeth A. Phelps. *The Cognitive Neuroscience of Mind: A Tribute to Michael S. Gazzaniga*. MIT Press, 2010. 1

[28] Pritam Sarkar and Ali Etemad. XKD: Cross-modal knowledge distillation with domain alignment for video representation learning. In *Proc. AAAI Conf. Artif. Intell.*, pages 14875–14885, 2024. 2

[29] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proc. Eur. Conf. Comput. Vis.*, pages 436–454, 2020. 2

[30] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proc. Eur. Conf. Comput. Vis.*, pages 252–268, 2018. 1, 2, 6

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 8

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, page 6000–6010, Red Hook, NY, USA, 2017. 4

[33] Hao Wang, Zheng-Jun Zha, Liang Li, Xuejin Chen, and Jiebo Luo. Semantic and relation modulation for audio-visual event localization. *IEEE Trans. Pattern. Anal. Mach. Intell.*, 45(6):7711–7725, 2023. 1, 2

[34] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 2

[35] Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Lilian Tang, Lin Li, Yongqiang Wang, Aoxiong Yin, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 22099–22114, 2023. 2

[36] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in Audio-visual Context: A Review, Analysis, and New Perspective. *arXiv e-prints*, page arXiv:2208.09579, 2022. 1

[37] Peng Wu, Xiaotao Liu, and Jing Liu. Weakly supervised audio-visual violence detection. *IEEE Trans. Multimedia*, 25:1674–1685, 2023. 1

[38] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. Int. Conf. Acoust. Speech, Signal Processing.*, pages 1–5. IEEE, 2023. 2, 3, 4, 6, 8

[39] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 6291–6299, 2019. 2, 6

[40] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 19957–19966, 2022. 1, 2, 5, 6, 7

[41] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proc. 28th ACM Int. Conf. Multimedia*, pages 3893–3901, 2020. 1, 2, 5, 6, 7

[42] Hanyu Xuan, Zhiliang Wu, Jian Yang, Bo Jiang, Lei Luo, Xavier Alameda-Pineda, and Yan Yan. Robust audio-visual contrastive learning for proposal-based self-supervised sound source localization in videos. *IEEE Trans. Pattern. Anal. Mach. Intell.*, pages 1–12, 2024. 2

[43] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proc. AAAI Conf. Artif. Intell.*, volume 34, pages 279–286, 2020. 2, 6

[44] Cheng Xue, Xionghu Zhong, Minjie Cai, Hao Chen, and Wenwu Wang. Audio-visual event localization by learning spatial and semantic co-attention. *IEEE Trans. Multimedia*, 25:418–429, 2023. 2, 6

[45] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 11542–11552, 2021. 2, 3

[46] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 9929–9938, 2020. 2

[47] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. MM-Pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proc. 30th ACM Int. Conf. Multimedia*, pages 6241–6249, 2022. 1

[48] Yihan Zhao, Wei Xi, Gairui Bai, Xinhui Liu, and Jizhong Zhao. Heterogeneous interactive graph network for audio–visual question answering. *Knowledge-Based Systems*, 300:112165, 2024. 4

[49] Jinxing Zhou, Dan Guo, Yiran Zhong, and Meng Wang. Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling. *International Journal of Computer Vision*, pages 1–22, 2024. 3

[50] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, pages 8432–8440, 2021. 1, 2, 6

[51] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *Int. J. Autom. Comput.*, 18(3):351–376, 2021. 1