Dense Audio-Visual Event Localization under Cross-Modal Consistency and Multi-Temporal Granularity Collaboration

Ziheng Zhou[1], Jinxing Zhou[1]*, Wei Qian[1], Shengeng Tang[1], Xiaojun Chang[2,3], Dan Guo[1]*

[1]School of Computer Science and Information Engineering, Hefei University of Technology
[2]University of Science and Technology of China
[3]Mohamed Bin Zayed University of Artificial Intelligence
zzhhfut@gmail.com, zhoujxhfut@gmail.com, guodan@hfut.edu.cn

Abstract

In the field of audio-visual learning, most research tasks focus exclusively on short videos. This paper focuses on the more practical Dense Audio-Visual Event Localization (DAVEL) task, advancing audio-visual scene understanding for longer, untrimmed videos. This task seeks to identify and temporally pinpoint all events simultaneously occurring in both audio and visual streams. Typically, each video encompasses dense events of multiple classes, which may overlap on the timeline, each exhibiting varied durations. Given these challenges, effectively exploiting the audio-visual relations and the temporal features encoded at various granularities becomes crucial. To address these challenges, we introduce a novel CCNet, comprising two core modules: the Cross-Modal Consistency Collaboration (CMCC) and the Multi-Temporal Granularity Collaboration (MTGC). Specifically, the CMCC module contains two branches: a cross-modal interaction branch and a temporal consistency-gated branch. The former branch facilitates the aggregation of consistent event semantics across modalities through the encoding of audio-visual relations, while the latter branch guides one modality's focus to pivotal event-relevant temporal areas as discerned in the other modality. The MTGC module includes a coarse-to-fine collaboration block and a fine-to-coarse collaboration block, providing bidirectional support among coarse- and fine-grained temporal features. Extensive experiments on the UnAV-100 dataset validate our module design, resulting in a new state-of-the-art performance in dense audio-visual event localization.

Introduction

Hearing and vision are two crucial senses for humans in perceiving their surroundings. Within the research community, recent years have seen a surge of interest in the joint exploration and comprehension of audio and visual signals, giving rise to numerous audio-visual learning tasks. These include audio-visual event localization (Tian et al. 2018b; Zhou et al. 2024a) and video parsing (Tian, Li, and Xu 2020; Zhou et al. 2023, 2024b,c; Gao, Chen, and Xu 2023; Zhao et al. 2024), sound source localization (Senocak et al. 2021; Hu, Nie, and Li 2019; Qian et al. 2020) and segmentation (Zhou et al. 2022, 2024d; Mao et al. 2023; Liu et al. 2023; Li et al. 2023; Guo et al. 2023), audio-visual question answering (Lao et al. 2023; Li et al. 2022; Yang et al. 2022; Li et al. 2024a,b; Li, Hou, and Hu 2023) and captioning (Tian et al. 2018a; Iashin and Rahtu 2020; Mao et al. 2024; Shen et al. 2023), *etc*. However, the majority of these tasks have predominantly focused on trimmed videos of short durations, commonly 5s or 10s.

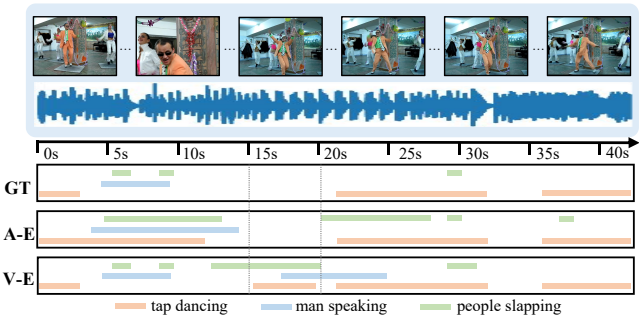

Figure 1: Illustration of the Dense Audio-Visual Event Localization (DAVEL) task. The DAVEL task requires temporally localizing the events that occur simultaneously in both audio and visual tracks of untrimmed videos. These dense events may overlap on the timeline and vary in duration. "GT" denotes the ground truth for audio-visual events, which are the intersection of audio events ("A-E") and visual events ("V-E").

In this paper, we focus on a more realistic task termed Dense Audio-Visual Event Localization (DAVEL) (Geng et al. 2023), which concentrates on the scene understanding of long, *untrimmed* audible videos. Notably, the DAVEL can be considered an extension of the existing audio-visual event localization (AVEL) (Tian et al. 2018b) task. Both tasks aim to identify and temporally localize the audio-visual events, events occurring in both audio and visual tracks. The key difference is that the AVEL focuses on short videos (fixed at 10s) where each video contains only a specific event class. In contrast, as illustrated in Fig. 1, the DAVEL task addresses *untrimmed* videos, with an average length of 42.1s for the official UnAV-100 (Geng et al. 2023) dataset. Furthermore, each video usually contains *dense* events, and events from multiple classes can temporally overlap, indicating the event's co-occurrence. More importantly, the AVEL is formulated as a *segment-level classification* problem, while the DAVEL seeks to precisely regress the start and end timestamps of each

*Corresponding authors

detected event (*frame-level regression*). This task difference renders prior excellent works in the AVEL task inapplicable to the DAVEL task. We will provide more introduction and discussions about this in the Related Work section.

Here, we highlight three main characteristics of the studied DAVEL task. **(C1) Cross-modal event-consistency.** The target audio-visual events are the *intersection* between audio events and visual events. In other words, a DAVEL model aims to capture event semantics *shared* between audio and visual modalities; **(C2) Cross-modal temporal consistency.** Not all temporal segments contain audio or visual events, indicating that some segments may harbor background noise or other event-irrelevant information. For instance, as depicted in Fig. 1, there are no *audio* events between 15s∼20s. As a result, the ground truth contains no *audio-visual* events regardless of the events present in the *visual* track during this temporal period. This decision is guided by the aforementioned intersection operation defined for audio-visual events. For one modality, the model should also focus on key temporal regions identified in the other modality; **(C3) Event duration inconsistency.** As presented in Fig. 1, each event may span various temporal windows. Considering interactions among audiovisual features at different temporal granularities would benefit the model.

Motivated by these observations, we propose a new CCNet for DAVEL task, comprising two core modules: the **Cross-Modal Consistency Collaboration (CMCC)** and the **Multi-Temporal Granularity Collaboration (MTGC)**. The details of each module are illustrated in Fig. 2. Specifically, the CMCC module draws its design from the analyses of the first two characteristics outlined above. The CMCC includes two branches: a cross-modal interaction branch and a temporal consistency-gated branch. The former branch encodes audio-visual relations through multi-head attention (Vaswani et al. 2017) mechanism, enabling each modality to aggregate *consistent event semantics* from the counterpart modality **(C1)**. The latter, a temporal consistency-gated branch, initiates by encoding unimodal relations within one modality via self-attention. Then, the encoded feature is utilized to learn a temporal weight vector highlighting *key event-related temporal regions*, subsequently serving as a temporal-wise gate to regularize the feature of the other modality **(C2)**. It is noteworthy that multiple stacked CMCC modules are utilized in our network, with a feature downsampling operation implemented at the outset of each CMCC module, producing features at multiple temporal scales. Regarding the MTGC module, it encompasses a Coarse-to-Fine collaboration (C2F) and a Fine-to-Coarse collaboration (F2C) block. In general, temporal features with relatively high downsampling rate are considered coarse-grained, while those with a lower downsampling rate are deemed fine-grained. Coarse-grained features prove advantageous in delineating coarse temporal regions of events occurring in the video, whereas fine-grained features aid in predicting precise temporal boundaries of events. The C2F and F2C blocks facilitate *bidirectional collaboration* between the coarse- and fine-grained features *across multiple temporal granularities*, benefiting the localization of events with varied durations **(C3)**.

Extensive experimental results on the UnAV-100 dataset demonstrate the effectiveness and superiority of our method. Our main contributions can be summarized as follows:

- We identify and analyze three key characteristics of the DAVEL task, which leads to a new CCNet approach comprising several simple yet highly effective modules.
- We design a Cross-Modal Consistency Collaboration module, which incorporates both a cross-modal interaction branch and a temporal consistency-gated branch, ensuring superior audio-visual representation embedding.
- We introduce a Multi-Temporal Granularity Collaboration module, which features coarse-to-fine and fine-to-coarse collaboration blocks, enabling the model to utilize temporal features bidirectionally across various granularities.
- Our method achieves a new state-of-the-art on the UnAV-100 dataset, surpassing the previous baseline in mAP metrics across multiple tIoU thresholds and exhibiting superior performance in localizing events of varied durations.

Related Work

Audio-Visual Event Localization (AVEL) task aims to identify the video segments containing a specific audio-visual event (both audible and visible) and classify its category. The pioneer work (Tian et al. 2018b) proposes a dual multimodal residual network to fuse audio and visual features. Zhou *et al*. (Zhou et al. 2021; Zhou, Guo, and Wang 2023) design a positive sample propagation module to select the most highly relevant audio-visual pairs for feature aggregation. To deal with audio-visual events existing in different temporal scales, Yu *et al*. (Yu et al. 2022) constrain the audio-visual interaction in multiple fixed-size temporal windows. Although those methods have achieved significant progress for the AVEL problem, they are designed for trimmed videos in a short duration and can only realize the *segment-level event classification*. In contrast, the studied DAVEL task tackles long untrimmed videos and requires *frame-level timestamp regression*, which can not be solved by those AVEL methods.

Audio-Visual Video Parsing (AVVP) task (Tian, Li, and Xu 2020) aims to comprehensively localize the audio events, visual events, and audio-visual events. Unlike the AVEL and the studied DAVEL tasks, AVVP task does not emphasize audio-visual alignment. Moreover, the AVVP task is performed in a weakly supervised setting, where only the event label of the whole video is available for model training. Some researchers focus on developing more effective methods for audio-visual feature interaction (Yu et al. 2022; Lin et al. 2021; Jiang et al. 2022; Zhou et al. 2024b). Others try to generate video-level (Wu and Yang 2021; Cheng et al. 2022) or segment-level pseudo labels (Yung-Hsuan Lai 2023; Zhou et al. 2024c) via label denoising or by utilizing pretrained large-scale models for better model optimization. However, these AVVP methods remain limited to short, trimmed videos, rendering them unsuitable for the DAVEL task.

Dense Audio-Visual Event Localization (DAVEL) task aims to temporally localize all the audio-visual events appearing in untrimmed videos, predicting the corresponding event categories and temporal boundaries. Notably, the DAVEL is a newly proposed research task. The pioneering work (Geng

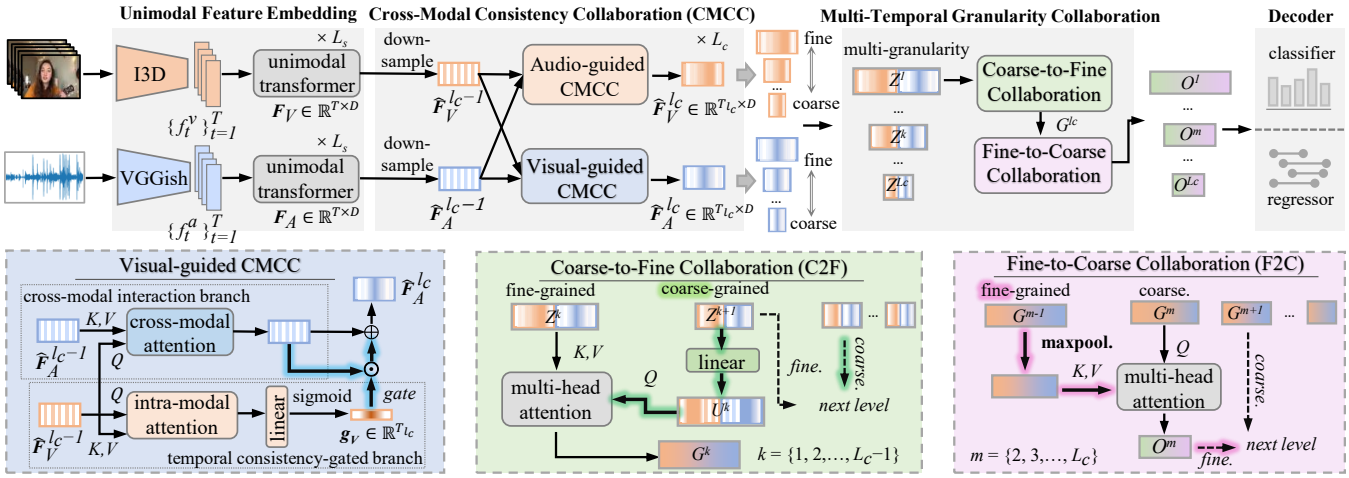

Figure 2: The pipeline of our CCNet for the dense audio-visual event localization task.

et al. 2023) serves as a strong baseline for comparison. Specifically, the backbone of this baseline includes a cross-modal pyramid transformer encoder (CMPT) and a temporal dependency modeling (TDM) module. The CMPT is used to encode the audio-visual relations at different temporal scales, while the TDM is designed to model the dependencies of the temporal segments and the concurrent events in different categories. Although this strong baseline has achieved impressive localization results, it simply encodes the cross-modal interaction via vanilla transformer encoders and overlooks the cooperation among features of various temporal scales. In contrast, our method considers enhancing the cross-modal consistency and multi-temporal granularity collaboration.

Methodology

Problem Formulation

Given an audible video sequence, it is first divided into T segments at equal intervals. Let $\{V_t, A_t\}_{t=1}^{T}$ denote the visual and audio segment pairs, respectively. The ground truth for each video is given as $Y = \{y_n = (t_{s,n}, t_{e,n}, c_n)\}_{n=1}^{N}$, indicating there are N audio-visual events in the video. For the n-th event, $t_{s,n}$ and $t_{e,n}$ are the start and end timestamps, respectively, and c_n represents the event category, where $c_n \in \{1, ..., C\}$ (C is the total class number of audio-visual events in the dataset).

The dense audio-visual event localization task is formulated as a sequence labeling and regression problem: for each timestamp t, the model needs to classify its event category and regress the distances from this moment to the event's start and end timestamps. Consequently, the prediction is given as $\hat{Y} = \{\hat{y}_t = (d_{s,t}, d_{e,t}, p(c_t))\}_{t=1}^{T}$, where $p(c_t) \in \mathbb{R}^{1\times C}$ is the event probability, $d_{s,t}$ and $d_{e,t}$ are the regressed onset and offset distances, respectively. It is noteworthy that $d_{s,t}$ and $d_{e,t}$ are predicted only when an audio-visual event exists at moment t. The final localization results can be obtained by post-processing the predictions as:

$$c_t = \text{argmax}(p(c_t)), \quad t_{s,t} = t - d_{s,t}, \quad t_{e,t} = t + d_{e,t}. \quad (1)$$

Framework Overview

The overall pipeline of our framework is illustrated in Fig. 2. It consists of four main modules: (1) At the initial **Unimodal Feature Embedding** module, video frames and audio signals are preprocessed, and their corresponding features are extracted using off-the-shelf pretrained convolutional neural networks. The audio and visual features are further refined by encoding unimodal temporal relations. (2) Subsequently, the proposed **Cross-Modal Consistency Collaboration (CMCC)** module generates audio and visual features at various temporal scales/granularities. Each CMCC block consists of *a cross-modal interaction branch* and *a temporal consistency-gated branch*. The former branch focuses on encoding cross-modal audio-visual relations to aggregate information on shared events in both modalities, while the latter branch ensures temporal consistency across audio and visual modalities, guiding the feature from one modality to focus on key temporal regions identified in the other modality. (3) The updated audio and visual features are concatenated and transferred to the **Multi-Temporal Granularity Collaboration (MTGC)** module, designed to bolster interactions among features across different temporal granularities. The MTGC contains *a Coarse-to-Fine (C2F)* and *a Fine-to-Coarse (F2C) collaboration* block, allowing a bidirectional feature flow between coarse-grained and fine-grained temporal features. (4) Finally, the refined audiovisual features are forwarded to the **Decoder**, which uses a classification head to predict event probabilities and a regression head to determine the onset and offset distances.

Unimodal Feature Embedding

Given audio and visual components $\{A_t, V_t\}_{t=1}^{T}$, we extract their features following the baseline (Geng et al. 2023). Specifically, the VGGish (Hershey et al. 2017) is utilized to extract the audio feature $\boldsymbol{a} = \{f_t^a\}_{t=1}^{T} \in \mathbb{R}^{T\times d_a}$. For the visual feature extraction, the two-stream I3D (Carreira and Zisserman 2017) model is used, yielding the visual feature $\boldsymbol{v} = \{f_t^v\}_{t=1}^{T} \in \mathbb{R}^{T\times d_v}$. Then, we use two convolutional layers to project the audio and visual features into the same

embedding space, yielding $\boldsymbol{a}, \boldsymbol{v} \in \mathbb{R}^{T\times D}$, where D is the feature dimension. Note that the lengths of videos may vary, thus the feature vectors obtained by the pretrained models are cropped or padded to the max sequence length T.

Considering that audio or visual signals representing an event typically occur in consecutive timestamps, we further encode the uni-modal temporal relations via the self-attention mechanism. This can be conveniently implemented by feeding the audio feature $\boldsymbol{a}$ or visual feature $\boldsymbol{v}$ into L_s stacked Transformer (Vaswani et al. 2017) blocks, with distinct parameters for each modality.

Through the aforementioned steps, we can obtain the unimodal audio and visual embeddings, denoted as $\boldsymbol{F}_A, \boldsymbol{F}_V \in \mathbb{R}^{T\times D}$, respectively. Next, we consider facilitating the dense audio-visual event localization from two task-relevant perspectives: cross-modal consistency and multi-temporal granularity collaboration. We elaborate on their design principles and detailed operations next.

Cross-modal Consistency Collaboration

The target audio-visual events in the video are both audible and visible. Except for encoding the unimodal relations, one modality should also be aware of events within another modality and aggregate consistent event information from the other. Given the audio and visual embeddings $\boldsymbol{F}_A, \boldsymbol{F}_V \in \mathbb{R}^{T\times D}$, we utilize **a Cross-Modal Interaction branch (CMI)** which encodes cross-modal relations through the multi-head attention (MHA) (Vaswani et al. 2017), followed by residual connection. Specifically, the feature of one modality is used as the *key* $\mathcal{K}$ and *value* $\mathcal{V}$, and the feature of another modality is used as the *query* $\mathcal{Q}$, formulated as,

$$\begin{aligned}\hat{\boldsymbol{F}}_A &= \boldsymbol{F}_A + \text{MHA}(\boldsymbol{F}_V, \boldsymbol{F}_A, \boldsymbol{F}_A),\\ \hat{\boldsymbol{F}}_V &= \boldsymbol{F}_V + \text{MHA}(\boldsymbol{F}_A, \boldsymbol{F}_V, \boldsymbol{F}_V),\\ \text{MHA}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) &= \delta(\frac{\mathcal{Q}\boldsymbol{W}^Q(\mathcal{K}\boldsymbol{W}^K)^\top}{\sqrt{D}})\mathcal{V}\boldsymbol{W}^V,\end{aligned} \tag{2}$$

where $\hat{\boldsymbol{F}}_{\boldsymbol{A}}, \hat{\boldsymbol{F}}_{\boldsymbol{V}} \in \mathbb{R}^{T\times D}$ are the updated audio and visual features, $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{D\times D}$ are learnable parameters, δ is the softmax function. Here, we simplify the presentation by omitting the feed-forward layers attached to the MHA operations. The MHA mechanism allows each audio (visual) segment to engage with all the visual (audio) segments. The features of one modality can be augmented by incorporating relevant event information from the other modality through high cross-modal attention weights (*event consistency*).

Furthermore, we consider that each modality may contain event-unrelated background noise within specific temporal segments. Each modality should focus on the key temporal regions of the counterpart modality that contain informative foreground. This is because a video segment contains an audio-visual event only if both the audio segment and visual segment include this event (*temporal consistency*). Therefore, we design **a Temporal Consistency-Gated branch (TCG)**. Specifically, the feature of one modality is first enhanced by intra-modal attention, then it is processed through a linear layer followed by the sigmoid activation (σ). This yields the temporal weight $\boldsymbol{g} \in \mathbb{R}^T$, which then serves as a consistency gate for the feature of the other modality. These operations are formulated as,

$$\begin{aligned}\boldsymbol{g}_V &= \sigma((\text{MHA}(\boldsymbol{F}_V, \boldsymbol{F}_V, \boldsymbol{F}_V))\boldsymbol{W}_v),\\ \hat{\boldsymbol{F}}_A &\Longleftarrow \hat{\boldsymbol{F}}_A + \boldsymbol{g}_V \odot \hat{\boldsymbol{F}}_A,\\ \boldsymbol{g}_A &= \sigma((\text{MHA}(\boldsymbol{F}_A, \boldsymbol{F}_A, \boldsymbol{F}_A))\boldsymbol{W}_a),\\ \hat{\boldsymbol{F}}_V &\Longleftarrow \hat{\boldsymbol{F}}_V + \boldsymbol{g}_A \odot \hat{\boldsymbol{F}}_V,\end{aligned} \tag{3}$$

where $\boldsymbol{W}_v, \boldsymbol{W}_a \in \mathbb{R}^{D\times 1}$ are learnable parameters of the linear layers, $\odot$ denotes element-wise multiplication, $\hat{\boldsymbol{F}}_A, \hat{\boldsymbol{F}}_V \in \mathbb{R}^{T\times D}$. Note that the latent feature $\boldsymbol{g}$ is automatically learned during model training. Under this guidance, the feature of each modality is further enhanced by focusing on the informative temporal regions recognized in the other modality.

For convenience, we denote the above operations of two branches in Eqs. 2 and 3 as a Cross-Modal Consistency Collaboration (CMCC) layer, symbolized as,

$$\hat{\boldsymbol{F}}_A, \hat{\boldsymbol{F}}_V = \text{CMCC}(\boldsymbol{F}_A, \boldsymbol{F}_V). \tag{4}$$

Inspired by the baseline (Geng et al. 2023), we incorporate cross-modal collaboration at different temporal scales to better perceive events of varied durations. Specifically, we utilize L_c stacked CMCC layers. Within each layer, the audio and visual features are first downsampled with a stride 2^{l_c-1}, where l_c is the index of the current layer. The downsampled features are used as the *query*, *key*, and *value* in the MHA operation (Eq. 2). The output of the l_c-1 CMCC layer is then utilized as the input to the l_c layer. Therefore, we obtain:

$$\hat{\boldsymbol{F}}_A^{l_c}, \hat{\boldsymbol{F}}_V^{l_c} = \text{CMCC}(\hat{\boldsymbol{F}}_A^{l_c-1}, \hat{\boldsymbol{F}}_V^{l_c-1}), \tag{5}$$

where $l_c \in \{1, 2, ..., L_c\}$, $\hat{\boldsymbol{F}}_A^0 = \boldsymbol{F}_A, \hat{\boldsymbol{F}}_V^0 = \boldsymbol{F}_V$, and $\hat{\boldsymbol{F}}_A^{l_c}, \hat{\boldsymbol{F}}_V^{l_c} \in \mathbb{R}^{T_{l_c}\times D}$ ($T_{l_c} = T/2^{l_c-1}$). Then, the audio and visual features at the same temporal scale are concatenated, resulting in the feature pyramid $\boldsymbol{Z} = \{\boldsymbol{Z}^{l_c}\}_{l_c=1}^{L_c}$, where $\boldsymbol{Z}^{l_c} = \text{Concat}(\hat{\boldsymbol{F}}_A^{l_c}, \hat{\boldsymbol{F}}_V^{l_c}) \in \mathbb{R}^{T_{l_c}\times 2D}$.

Multi-Temporal Granularity Collaboration

After obtaining the audiovisual features at different temporal scales, the previous baseline (Geng et al. 2023) directly sends them into a Temporal Dependency Modelling (TDM) module, which models the dependencies of simultaneous events and consecutive segments. However, the TDM operates on features at separate temporal scales. In contrast, we propose a very simple but effective Multi-Temporal Granularity Collaboration (MTGC) module to enhance the collaboration among different temporal scales. Our motivation is that the *coarse* temporal features ($\boldsymbol{Z}^{l_c}$ with larger l_c) have a larger receptive field for recognizing the occurring event, while the *fine*-grained features ($\boldsymbol{Z}^{l_c}$ with smaller l_c) are more beneficial for precise event boundary prediction. Our proposed MTGC module explores a bidirectional collaboration mechanism for better temporal localization, as detailed next.

Coarse-to-Fine Collaboration (C2F). Given the concatenated audiovisual features $\boldsymbol{Z}^k \in \mathbb{R}^{T_k\times 2D}$ ($T_k = T/2^{k-1}$) at the k-th temporal granularity ($k \in \{1, 2, ..., L_c - 1\}$), we treat it as the current *fine*-grained temporal feature. Then, the feature at adjacent $k+1$ granularity $\boldsymbol{Z}^{k+1}$ can be regarded

as the *coarse*-grained feature. We then apply a linear layer followed by the ReLU activation to transform the coarse-grained feature to match the dimension of the fine-grained feature $\boldsymbol{Z}^k$, written as,

$$\boldsymbol{U}^k = \text{ReLU}(\boldsymbol{W}_u \boldsymbol{Z}^{k+1}), \tag{6}$$

where $\boldsymbol{W}_u \in \mathbb{R}^{T_k \times T_{k+1}}$ is the learnable paramter of the linear layer, $\boldsymbol{U}^k \in \mathbb{R}^{T_k \times 2D}$. In principle, $\boldsymbol{U}^k$ provides the event information from a more coarse-grained level, which can collaborate with the fine-grained $\boldsymbol{Z}^k$. We achieve the coarse-to-fine collaboration between $\boldsymbol{U}^k$ and $\boldsymbol{Z}^k$ via simple multi-head attention (MHA), thus generating the updated feature at k-th temporal granularity $\boldsymbol{G}^k$:

$$\boldsymbol{G}^k = \text{MHA}(\boldsymbol{U}^k, \boldsymbol{Z}^k, \boldsymbol{Z}^k), \tag{7}$$

where $\boldsymbol{G}^k \in \mathbb{R}^{T_k \times D}, k = \{1, 2, ..., L_c - 1\}$. Notably, for the largest L_c-th granularity, there are no more coarse-grained temporal features, so we set $\boldsymbol{G}^{L_c} = \boldsymbol{Z}^{L_c}$. After the C2F collaboration among temporal features at different granularities, we adopt the TDM (Geng et al. 2023) module to further enhance features at each separate temporal scale.

Fine-to-Coarse Collaboration (F2C). In addition to the coarse-to-fine collaboration, we also develop a fine-to-coarse collaboration mechanism that enables a bidirectional interaction for features at multiple temporal granularities. Assuming the updated feature $\boldsymbol{G}^m \in \mathbb{R}^{T/2^{m-1} \times 2D}$ at the m-th temporal granularity ($m \in \{2, ..., L_c\}$) as *coarse*-grained feature, the feature $\boldsymbol{G}^{m-1} \in \mathbb{R}^{T/2^{m-2} \times 2D}$ at the adjacent $m-1$ temporal granularity can be regarded as the *fine*-grained feature.

We first temporally downsample the fine-grained feature $\boldsymbol{G}^{m-1}$ via max-pooling to align its dimension with $\boldsymbol{G}^m$. Then, we model the fine-to-coarse collaboration using multi-head attention. These operations can be summarized as,

$$\begin{aligned} \hat{\boldsymbol{G}}^{m-1} &= \text{MaxPooling}(\boldsymbol{G}^{m-1}), \\ \boldsymbol{O}^m &= \boldsymbol{G}^m + \text{MHA}(\boldsymbol{G}^m, \hat{\boldsymbol{G}}^{m-1}, \hat{\boldsymbol{G}}^{m-1}), \end{aligned} \tag{8}$$

where $\boldsymbol{O}^m \in \mathbb{R}^{T/2^{m-1} \times 2D}$ is the updated feature at the m-th temporal granularity, $m = 2, ..., L_c$. Let $\boldsymbol{O}^m = \text{F2C}(\boldsymbol{G}^{m-1}, \boldsymbol{G}^m)$ represent the fine-to-coarse collaboration process described above, the output $\boldsymbol{O}^m$ is used as the *fine*-grained feature in the next m+1 granularity: $\boldsymbol{O}^{m+1} = \text{F2C}(\boldsymbol{O}^m, \boldsymbol{G}^{m+1})$. For the temporal granularity $m = 1$, there are no more fine-grained features, we simply assign $\boldsymbol{O}^1 = \boldsymbol{G}^1$. In this way, the temporal features at each granularity $\boldsymbol{O}^{l_c} (l_c = 1, 2, ..., L_c)$ are enhanced by incorporating both coarse- and fine-grained event clues, which are ready for decoding audio-visual events across varied temporal ranges.

Decoder

Following the paradigm of baseline (Geng et al. 2023), the decoder of the DAVEL task includes a classification head and a regression head. Given the feature $\boldsymbol{O}^{l_c} \in \mathbb{R}^{T/2^{l_c-1} \times 2D}$ at the l_c temporal granularity ($l_c = \{1, 2, ..., L_c\}$), the classification head predicts the corresponding event probability $p(c_t)$ for each timestamp t. The classification head is implemented by three 1D convolution layers following a sigmoid function. As for the regression head, it also consists of three 1D convolutions but is activated with the ReLU function. This head directly regresses the distances from the current timestamp t to the start and end timestamp of an event $(d_{s,t}, d_{e,t})$ if the event exists. The regression output with the shape of $[2, C, T_{l_c}]$ indicates the onsets and offsets to an event at each timestamp, which is also class-aware for recognizing overlapping events with different categories.

Training. We train our model by employing two losses, *i.e.*, a focal loss (Lin et al. 2017) $\mathcal{L}_{cls}$ for imbalanced event classification, and a generalized IoU loss (Rezatofighi et al. 2019) $\mathcal{L}_{reg}$ for distance regression. The total training objective can be written as,

$$\mathcal{L} = \alpha \sum_t \mathcal{L}_{cls} + \beta \sum_t \mathbb{I}_t \mathcal{L}_{reg}, \tag{9}$$

where α and β are two hyperparameters, which are identical to those in baseline (Geng et al. 2023), $\mathbb{I}_t$ is a function indicating whether a timestamp t contains audio-visual events.

Inference. For each timestamp, we predict its event classes and the temporal boundary of each event following Eq. 1. The results are then post-processed using the Soft-NMS (Bodla et al. 2017) technique to suppress predictions that are predicted to be in the same category but are highly overlapping.

Experiments

Experimental Setups

Dataset. Our experiments are conducted on the official UnAV-100 (Geng et al. 2023) dataset, specifically constructed for the dense audio-visual event localization task. 1) *Untrimmed videos.* The UnAV-100 dataset comprises 10,790 untrimmed videos with varied temporal lengths, with the majority exceeding 40 seconds. 2) *Multiple categories.* The videos encompass 100 categories of audio-visual events commonly found in natural environments, including human or animal activities, musical instruments, various vehicles, *etc*. *3) Densely overlapping events of varying durations.* On average, each video features 2.8 audio-visual events, highlighting the presence of multiple overlapping events. Furthermore, these events span a range of durations (*i.e.*, distinct temporal windows). These characteristics make event localization challenging. Following the standard dataset split, the distribution among training, validation, and testing subsets is set at a ratio of 3:1:1.

Evaluation metric. Following the baseline (Geng et al. 2023), we adopt the mean Average Precision (mAP) as the metric for evaluating temporal localization results. We report mAPs at tIoU thresholds ranging from 0.5 to 0.9 in increments of 0.1 ([0.5:0.1:0.9]). Additionally, we report the average mAP (denoted as 'Avg.'), calculated across an expanded range of thresholds [0.1:0.1:0.9], serving as a comprehensive measure for comparing overall model performance.

Implementation details. For visual feature extraction, frames are sampled at 25 FPS for each video, and the RAFT (Teed and Deng 2020) is utilized to extract the optical flow. Then, 24 consecutive RGB and the optimal flow frames are sent into the two-stream pretrained I3D (Carreira and Zisserman 2017) model, yielding 2048-D visual features. For audio feature extraction, audio signals are first split every 0.96s using a sliding window of 0.32s. Then, the pretrained

Methods	0.5	0.6	0.7	0.8	0.9	Avg.
VSGN (Zhao, Thabet, and Ghanem 2021)	24.5	20.2	15.9	11.4	6.8	24.1
TadTR (Liu et al. 2022)	30.4	27.1	23.3	19.4	14.3	29.4
ActionFormer (Zhang, Wu, and Li 2022)	43.5	39.4	33.4	27.3	17.9	42.2
DAVEL (Geng et al. 2023)	50.6	45.8	39.8	32.4	21.1	47.8
CCNet (ours)	**51.9**	**47.2**	**41.5**	**34.1**	**23.0**	**49.2**
$^{\triangle}$DAVEL (Geng et al. 2023)	53.8	48.7	42.2	33.8	20.4	51.0
$^{\triangle}$UniAV (Geng et al. 2024)	54.8	49.4	43.2	35.3	22.5	51.7
$^{\triangle}$**CCNet (ours)**	**57.3**	**52.2**	**46.2**	**38.1**	**25.6**	**54.1**

Table 1: Comparison with prior works. '$^{\triangle}$' denotes that more advanced audio and visual features extracted by ONE-PEACE (Wang et al. 2023) are used.

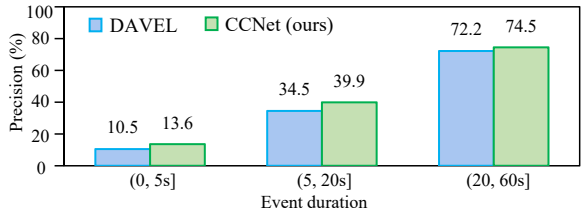

Figure 3: Localization results for various durations.

VGGish (Hershey et al. 2017) is used to extract features for each audio segment, resulting in the 128-D audio features. It is noteworthy that the video sequences vary in length; thus, we crop or pad the extracted audio and visual features to a maximum length of T=224. The layer numbers L_s and L_c are empirically set to 2 and 6, respectively. We train our model for 40 epochs with a batch size of 16. The Adam optimizer is used, with the initial learning rate and the weight decay set to 1e-4. Experiments are conducted on a NVIDIA A40 GPU.

Comparison to Prior Works

To ensure a more comprehensive comparison, we also compare our method with other representative approaches tailored for the temporal localization task, namely VSGN (Zhao, Thabet, and Ghanem 2021), TadTR (Liu et al. 2022), and ActionFormer (Zhang, Wu, and Li 2022). Given that we focus on a multimodal task, these methods are adapted to utilize concatenated audio and visual features as inputs. As shown in Table 1, these temporal action localization models significantly lag behind the baseline DAVEL (Geng et al. 2023). This underscores the critical importance of specialized designs for cross-modal relation modeling in the dense audio-visual event localization task. Furthermore, our method is superior to the baseline DAVEL (Geng et al. 2023) method. We attain a new state-of-the-art performance, achieving an average mAP of 49.2%. Our method surpasses the baseline DAVEL across all tIoUs thresholds, with notable improvements of 1.9% at the stringent tIoU=0.9. Furthermore, our method can be significantly enhanced by adopting more advanced audio-visual features extracted using ONE-PEACE (Wang et al. 2023), continuing to surpass previous baselines. These results demonstrate the superiority of our method, attributed to the proposed two core modules that facilitate cross-modal consistency and multi-temporal granularity collaborations.

In addition, we compare the proposed method with the baseline DAVEL (Geng et al. 2023) on the localization of events in different temporal durations. We analyze the events in videos from the UnAV-100 dataset, focusing on the event durations at $(0s, 5s]$ (short), $(5s, 20s]$ (middle), and $(20s, 60s]$ (long). For one event in a specific duration, we consider this event correctly localized if the tIoU between the model prediction and the ground truth exceeds a threshold of 0.5 and the predicted event category is correct. Then, we calculate the percentage of the correct localized events relative to the total number of events within that duration (*precision*). As shown in Fig. 3, our method surpasses the baseline in event localization across varied temporal durations. Particularly, our method achieves a 5.4% improvement over the baseline for events with a duration of $(5s, 20s]$. These improvements can be attributed to the multi-temporal granularity collaboration module in our method, which enables the effective integration of temporal cues across various scales.

CMCC	MTGC		0.5	0.6	0.7	0.8	0.9	Avg.
	C2F	F2C						
✔	✘	✘	50.5	45.6	39.8	33.1	22.9	47.9
✔	✔	✘	51.3	46.1	40.0	32.8	22.4	48.3
✔	✘	✔	50.8	46.2	40.6	34.0	**23.4**	48.2
✔	✔	✔	**51.9**	**47.2**	**41.5**	**34.1**	23.0	**49.2**

Table 2: The ablation study of our core modules.

Ablation Studies

Effectiveness of our core modules. Our method consists of the core Cross-Modal Consistency Collaboration (CMCC) and Multi-Temporal Granularity Collaboration (MTGC) modules. Specifically, the MTGC module encompasses the C2F and F2C collaboration blocks. We conduct ablation experiments to explore their impacts. As presented in Table 2, the average mAP is 47.9% when utilizing only the CMCC module. Note that this variant model still marginally outperforms the baseline DAVEL model regarding the average mAP. Moreover, this variant model exceeds DAVEL by 1.8% in precision at the tIoU threshold of 0.9, highlighting our superior localization capabilities. While DAVEL also considers encoding cross-modal relations, our CMCC module additionally incorporates temporal consistency within audio and visual modalities. The model performance can be improved by separately adding the C2F or the F2C block, indicating the effectiveness of each collaboration mechanism. Ultimately, our model achieves the highest performance when employing both C2F and F2C blocks simultaneously. This suggests that the bidirectional C2F and F2C work collaboratively, contributing to better exploitation from multi-temporal granularities.

Ablation study of the CMCC module. We assess the effectiveness of each branch of the CMCC module: the Cross-Modal Interaction (CMI) branch and the Temporal Consistency-Gated (TCG) branch. The experimental results are shown in the lower part of Table 3. We find that each branch is beneficial for improving event localization performance. The CMI branch facilitates one modality in aggregating relevant or complementary event semantics from the

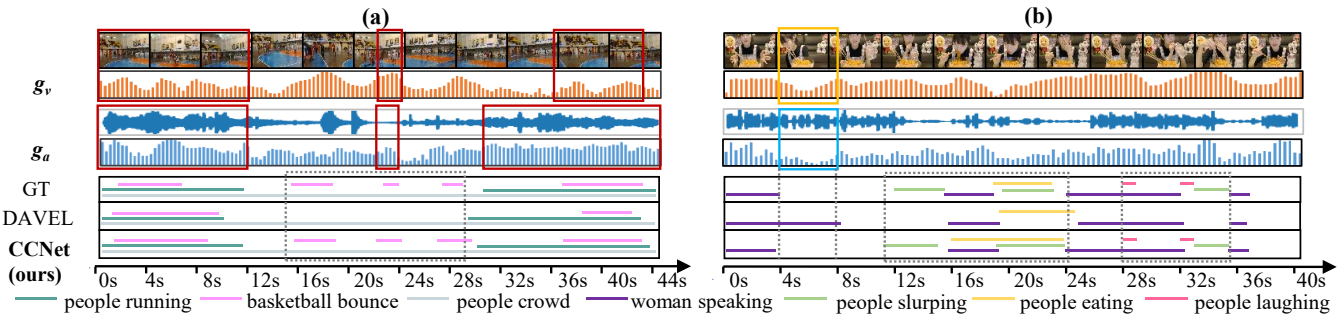

Figure 4: Qualitative examples of dense audio-visual event localization.

CMI	TCG	0.5	0.6	0.7	0.8	0.9	Avg.
✔	✔	**51.9**	**47.2**	**41.5**	**34.1**	**23.0**	**49.2**
✔	✘	50.8	46.1	40.6	33.6	22.6	48.1
✘	✔	49.9	45.3	40.0	33.2	22.8	47.6

Table 3: The ablation results of our CMCC module. We ablate the two branches in CMCC: the abbreviation 'CMI' is short for the Cross-Modal Interaction branch, while 'TCG' represents the Temporal Consistency-Gated branch.

Strategies	0.5	0.6	0.7	0.8	0.9	Avg.
C2F → F2C	**51.9**	**47.2**	**41.5**	**34.1**	**23.0**	**49.2**
F2C → C2F	49.9	45.3	40.1	32.7	21.8	47.1

Table 4: Ablation study on the operational sequence of C2F and F2C collaboration blocks in MTGC module. 'A' → 'B' denotes the initial application of block 'A', followed by the execution of block 'B' in MTGC module implementation.

other modality; while the TCG branch enables one modality to recognize the key temporal regions in the other modality. Consequently, the model performance can be enhanced by utilizing these two branches simultaneously.

Ablation study of the MTGC module. In the proposed MTGC module, the Coarse-to-Fine (C2F) block is implemented first, followed by the Fine-to-Coarse (F2C) collaboration block. Here, we explore the impacts of the operational sequence of these two blocks. The results, as shown in Table 4, indicate that applying C2F before F2C results in a higher average mAP of 49.2%, whereas the reversed order leads to a lower performance of 47.1%. This suggests that the order in which the C2F and F2C blocks are applied influences the overall performance of the MTGC module. In our supplementary material, we provide additional ablation studies on the proposed CMCC and MTGC modules.

Qualitative Results

We present some qualitative examples of dense audio-visual event localization. Fig. 4 (a) illustrates a video sample containing three classes of audio-visual events: *people running*, *basketball bounce*, and *people crowd*, densely distributed across various temporal extents. Compared to the baseline model DAVEL (Geng et al. 2023), our method demonstrates superior performance in temporally localizing events with varying durations. For instance, DAVEL fails to recognize event *basketball bounce* within the 15s∼30s (marked by the gray dotted box). In contrast, our method not only successfully identifies this event but also provides satisfactory temporal boundaries, highlighting the advantages of our proposed multi-temporal granularity collaboration mechanism. We also plot the curves of the learned temporal consistency gates $\boldsymbol{g}_A$ and $\boldsymbol{g}_V$, which indeed assign higher weights to temporal regions associated with these events (highlighted by the red boxes in the figure). In Fig. 4(b), the baseline DAVEL overlooks the audio-visual events *people slurping* and *people laughing* (gray dotted box). Conversely, our method accurately predicts the event categories and determines precise start and end timestamps. Furthermore, DAVEL incorrectly identifies an audio-visual event, *woman speaking*, within the 4s∼8s. However, our model's temporal consistency gate $\boldsymbol{g}_A$ assigns very low weights to this period (blue box), indicating the absence of audio events. Despite the clear depiction of *woman speaking* in the visual frames, as indicated by the high $\boldsymbol{g}_V$ values during this period (yellow box), the lack of corresponding audio events evidenced by our model ensures accurate prediction for this audio-visual event. These results confirm the effectiveness of the proposed cross-modal consistency collaboration mechanism.

Conclusion

We tackle a practical task of dense audio-visual event localization, which aims to temporally localize the audio-visual events densely occurring in untrimmed audible videos. We introduce a new CCNet approach and formulate its core modules with consideration of two essential perspectives: Cross-Modal Consistency Collaboration (CMCC) and Multi-Temporal Granularity Collaboration (MTGC). The CMCC module utilizes a cross-modal interaction branch to encode audio-visual interactions and incorporates a temporal consistency-gated branch to regulate each modality's focus on event-related temporal regions. The MTGC module consists of a coarse-to-fine and a fine-to-coarse collaboration block, which are beneficial for bidirectional cooperation among coarse- and fine-grained features across various temporal granularities. Experimental results demonstrate that our method surpasses previous baselines in accurately localizing dense audio-visual events of varying durations.

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers for their invaluable comments and insightful suggestions. The computation is completed on the HPC Platform of Hefei University of Technology. This work was supported by the National Natural Science Foundation of China (62272144), the Major Project of Anhui Province (2408085J040), and the Fundamental Research Funds for the Central Universities (JZ2024HGTG0309, JZ2024AHST0337).

References

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS — Improving Object Detection with One Line of Code. In *ICCV*, 5562–5570.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.

Cheng, H.; Liu, Z.; Zhou, H.; Qian, C.; Wu, W.; and Wang, L. 2022. Joint-Modal Label Denoising for Weakly-Supervised Audio-Visual Video Parsing. In *ECCV*, 431–448.

Gao, J.; Chen, M.; and Xu, C. 2023. Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception. In *CVPR*, 18827–18836.

Geng, T.; Wang, T.; Duan, J.; Cong, R.; and Zheng, F. 2023. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *CVPR*, 22942–22951.

Geng, T.; Wang, T.; Zhang, Y.; Duan, J.; Guan, W.; and Zheng, F. 2024. UniAV: Unified Audio-Visual Perception for Multi-Task Video Localization. *arXiv preprint arXiv:2404.03179*.

Guo, R.; Ying, X.; Chen, Y.; Niu, D.; Li, G.; Qu, L.; Qi, Y.; Zhou, J.; Xing, B.; Yue, W.; Shi, J.; Wang, Q.; Zhang, P.; and Liang, B. 2023. Audio-Visual Instance Segmentation. *arXiv preprint arXiv:2310.18709*.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP*, 131–135.

Hu, D.; Nie, F.; and Li, X. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 9248–9257.

Iashin, V.; and Rahtu, E. 2020. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*.

Jiang, X.; Xu, X.; Chen, Z.; Zhang, J.; Song, J.; Shen, F.; Lu, H.; and Shen, H. T. 2022. DHHN: Dual Hierarchical Hybrid Network for Weakly-Supervised Audio-Visual Video Parsing. In *ACM MM*, 719–727.

Lao, M.; Pu, N.; Liu, Y.; He, K.; Bakker, E. M.; and Lew, M. S. 2023. COCA: COllaborative CAusal Regularization for Audio-Visual Question Answering. In *AAAI*, 12995–13003.

Li, G.; Hou, W.; and Hu, D. 2023. Progressive Spatio-temporal Perception for Audio-Visual Question Answering. In *ACM MM*, 7808–7816.

Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.-R.; and Hu, D. 2022. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *CVPR*, 19108–19118.

Li, K.; Yang, Z.; Chen, L.; Yang, Y.; and Xiao, J. 2023. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM MM*, 1485–1494.

Li, Z.; Guo, D.; Zhou, J.; Zhang, J.; and Wang, M. 2024a. Object-Aware Adaptive-Positivity Learning for Audio-Visual Question Answering. In *AAAI*, 3306–3314.

Li, Z.; Zhou, J.; Zhang, J.; Tang, S.; Li, K.; and Guo, D. 2024b. Patch-level Sounding Object Tracking for Audio-Visual Question Answering. *arXiv preprint arXiv:2412.10749*.

Lin, T.-Y.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *ICCV*, 2999–3007.

Lin, Y.-B.; Tseng, H.-Y.; Lee, H.-Y.; Lin, Y.-Y.; and Yang, M.-H. 2021. Exploring Cross-Video and Cross-Modality Signals for Weakly-Supervised Audio-Visual Video Parsing. In *NeurIPS*.

Liu, C.; Li, P. P.; Qi, X.; Zhang, H.; Li, L.; Wang, D.; and Yu, X. 2023. Audio-Visual Segmentation by Exploring Cross-Modal Mutual Semantics. In *ACM MM*, 7590–7598.

Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; and Bai, X. 2022. End-to-end temporal action detection with transformer. *TIP*, 31: 5427–5441.

Mao, Y.; Shen, X.; Zhang, J.; Qin, Z.; Zhou, J.; Xiang, M.; Zhong, Y.; and Dai, Y. 2024. TAVGBench: Benchmarking text to audible-video generation. In *ACM MM*, 6607–6616.

Mao, Y.; Zhang, J.; Xiang, M.; Zhong, Y.; and Dai, Y. 2023. Multimodal Variational Auto-encoder based Audio-Visual Segmentation. In *ICCV*, 954–965.

Qian, R.; Hu, D.; Dinkel, H.; Wu, M.; Xu, N.; and Lin, W. 2020. Multiple Sound Sources Localization from Coarse to Fine. In *ECCV*, 1–16.

Rezatofighi, S. H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR*, 658–666.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and Kweon, I. S. 2021. Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications. In *TPAMI*, 1605–1619.

Shen, X.; Li, D.; Zhou, J.; Qin, Z.; He, B.; Han, X.; Li, A.; Dai, Y.; Kong, L.; Wang, M.; et al. 2023. Fine-grained audible video description. In *CVPR*, 10585–10596.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 402–419.

Tian, Y.; Guan, C.; Goodman, J.; Moore, M.; and Xu, C. 2018a. An attempt towards interpretable audio-visual video captioning. *arXiv preprint arXiv:1812.02872*.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 436–454.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018b. Audio-visual event localization in unconstrained videos. In *ECCV*, 247–263.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 1–15.

Wang, P.; Wang, S.; Lin, J.; Bai, S.; Zhou, X.; Zhou, J.; Wang, X.; and Zhou, C. 2023. ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities. *arXiv preprint arXiv:2305.11172*.

Wu, Y.; and Yang, Y. 2021. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, 1326–1335.

Yang, P.; Wang, X.; Duan, X.; Chen, H.; Hou, R.; Jin, C.; and Zhu, W. 2022. Avqa: A dataset for audio-visual question answering on videos. In *ACM MM*, 3480–3491.

Yu, J.; Cheng, Y.; Zhao, R.-W.; Feng, R.; and Zhang, Y. 2022. MM-Pyramid: Multimodal Pyramid Attentional Network for Audio-Visual Event Localization and Video Parsing. In *ACM MM*, 6241–6249.

Yung-Hsuan Lai, Y.-C. F. W., Yen-Chun Chen. 2023. Modality-Independent Teachers Meet Weakly-Supervised Audio-Visual Event Parser. In *NeurIPS*.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *ECCV*, 492–510.

Zhao, C.; Thabet, A. K.; and Ghanem, B. 2021. Video self-stitching graph network for temporal action localization. In *ICCV*, 13658–13667.

Zhao, P.; Zhou, J.; Guo, D.; Zhao, Y.; and Chen, Y. 2024. Multimodal Class-aware Semantic Enhancement Network for Audio-Visual Video Parsing. *arXiv preprint arXiv:2412.11248*.

Zhou, J.; Guo, D.; Guo, R.; Mao, Y.; Hu, J.; Zhong, Y.; Chang, X.; and Wang, M. 2024a. Towards Open-Vocabulary Audio-Visual Event Localization. *arXiv preprint arXiv:2411.11278*.

Zhou, J.; Guo, D.; Mao, Y.; Zhong, Y.; Chang, X.; and Wang, M. 2024b. Label-anticipated Event Disentanglement for Audio-Visual Video Parsing. In *ECCV*, 1–22.

Zhou, J.; Guo, D.; and Wang, M. 2023. Contrastive positive sample propagation along the audio-visual event line. *TPAMI*, 7239–7257.

Zhou, J.; Guo, D.; Zhong, Y.; and Wang, M. 2023. Improving audio-visual video parsing with pseudo visual labels. *arXiv preprint arXiv:2303.02344*.

Zhou, J.; Guo, D.; Zhong, Y.; and Wang, M. 2024c. Advancing Weakly-Supervised Audio-Visual Video Parsing via Segment-wise Pseudo Labeling. *IJCV*, 1–22.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2024d. Audio-Visual Segmentation with Semantics. *IJCV*, 1–21.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–visual segmentation. In *ECCV*, 386–403.

Zhou, J.; Zheng, L.; Zhong, Y.; Hao, S.; and Wang, M. 2021. Positive sample propagation along the audio-visual event line. In *CVPR*, 8436–8444.