Multimodal Class-aware Semantic Enhancement Network for Audio-Visual Video Parsing

Pengcheng Zhao*, Jinxing Zhou*, Yang Zhao, Dan Guo†, Yanxiang Chen†

School of Computer Science and Information Engineering, Hefei University of Technology
{stevenzhao1001, zhoujxhfut}@gmail.com, {yzhao, guodan, chenyx}@hfut.edu.cn

Abstract

The Audio-Visual Video Parsing task aims to recognize and temporally localize all events occurring in either the audio or visual stream, or both. Capturing accurate event semantics for each audio/visual segment is vital. Prior works directly utilize the extracted holistic audio and visual features for intra- and cross-modal temporal interactions. However, each segment may contain multiple events, resulting in semantically mixed holistic features that can lead to semantic interference during intra- or cross-modal interactions: the event semantics of one segment may incorporate semantics of unrelated events from other segments. To address this issue, our method begins with a Class-Aware Feature Decoupling (CAFD) module, which explicitly decouples the semantically mixed features into distinct class-wise features, including multiple event-specific features and a dedicated background feature. The decoupled class-wise features enable our model to selectively aggregate useful semantics for each segment from clearly matched classes contained in other segments, preventing semantic interference from irrelevant classes. Specifically, we further design a Fine-Grained Semantic Enhancement module for encoding intra- and cross-modal relations. It comprises a Segment-wise Event Co-occurrence Modeling (SECM) block and a Local-Global Semantic Fusion (LGSF) block. The SECM exploits inter-class dependencies of concurrent events within the same timestamp with the aid of a new event co-occurrence loss. The LGSF further enhances the event semantics of each segment by incorporating relevant semantics from more informative global video features. Extensive experiments validate the effectiveness of the proposed modules and loss functions, resulting in a new state-of-the-art parsing performance.

1 Introduction

In this paper, we focus on the task of Audio-Visual Video Parsing (AVVP) for the fundamental understanding of audio-visual scenes. As illustrated in Fig. 1 (a), given an audible video, the AVVP task seeks to temporally localize all events of interest, including audio events, visual events, and audio-visual events (both audible and visible). Notably, the audio and visual signals are not required to be temporally aligned in this task. As shown in Fig. 1 (a), the event *Speech* is present only in the audio modality while the *Accordion* is solely present in the visual modality. Moreover, multiple events can overlap on the timeline. To discern various types of events, an AVVP model needs to *accurately perceive event semantics for each audio/visual segment*. Given the multimodal and temporal characteristics of this task, the model must be carefully designed to not only utilize event semantics shared across audio and visual modalities but also maintain the unique event semantics inherent to each modality.

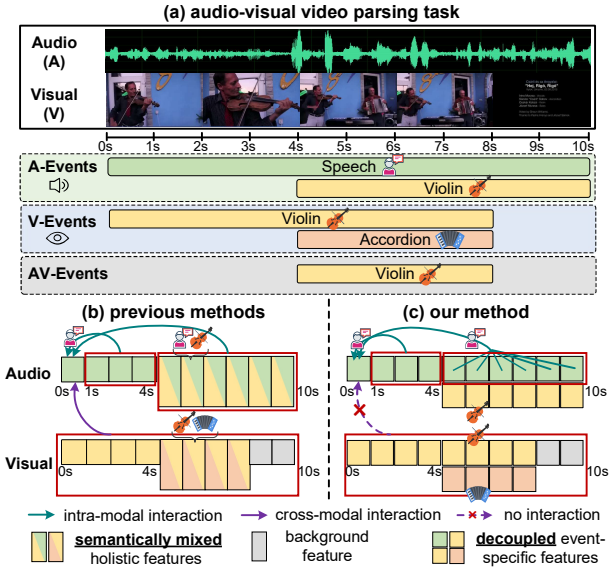

Figure 1: (a) Illustration of the AVVP task. (b) Previous methods rely on semantically mixed holistic audio/visual features for intra- and cross-modal interactions, leading to semantic interference. (c) In contrast, we utilize decoupled class-aware features to aggregate event semantics for each segment from only matched classes during interactions.

Previous works (Tian, Li, and Xu 2020; Yu et al. 2022; Lai, Chen, and Yu-Chiang 2023) have progressed this task by developing various sophisticated network architectures. For instance, one representative work, HAN (Tian, Li, and Xu 2020), introduces a hybrid attention network that utilizes self-attention and cross-attention mechanisms to capture event semantics through the modeling of temporal relations within

*These authors contributed equally.
†Corresponding authors.

intra- and cross-modalities, enabling each temporal segment to interact with all segments from the same or the other modality. Although these methods achieve impressive video parsing performances, they directly operate on *semantically mixed* audio and visual holistic features extracted by pre-trained audio/image classification backbones (Hershey et al. 2017; He et al. 2016), which could lead to **semantic interference** during intra- and cross-modal temporal modeling. Specifically, as illustrated in Fig. 1 (b), the first audio segment only contains the event 'Speech'. During intra-modal temporal modeling, interactions between this segment and those audio segments at (1s, 4s] will undoubtedly be beneficial for enhancing its event semantic of 'Speech' because they share the identical single 'Speech' event; however, this segment will incorporate the unrelated semantic of event 'Violin' when interacting with those audio segments at (4s, 10s], causing intra-modal semantic interference. Similarly, semantic interference can also occur during the cross-modal temporal modeling when the audio segment interacts with visual segments containing only partially similar or completely different events, *e.g.*, the interactions between the first audio segment and all visual segments in the example shown in Fig. 1 (b). To address this issue, we propose investigating *whether we can enhance the event semantics of each audio/visual segment by selectively aggregating only the most closely matched semantics from other segments during the intra-modal and cross-modal interactions*.

To this end, we first propose a **Class-Aware Feature Decoupling (CAFD)** module to explicitly decouple the semantically mixed holistic features of each audio/visual segment into distinct class-wise features, allowing our model to perform the interactions at a more fine-grained and precise class level. Specifically, the feature decoupling can be implemented by employing multiple independent linear layers on the holistic audio/visual feature for each segment. Particularly, in addition to the *event-specific* classes, we include a special *background* class in our feature decoupling process. This design is motivated by two observations: 1) each segment contains a degree of background information, and some background context may facilitate the event recognition. For example, the background of parking lot suggests that the event may relate to the cars; 2) some segments may contain only useless background noise, *e.g.*, the last two visual segments in Fig. 1 (a). Features of such segments should not be decoupled into other event-specific classes. Considering these observations, we employ a weighting mechanism to dynamically blend the background feature into the event-specific features to better utilize the decoupled background information. Notably, we introduce a reconstruction loss and an orthogonality loss to guarantee effective feature decoupling.

Then, we propose a **Fine-Grained Semantic Enhancement (FGSE)** module to fully exploit the decoupled class-aware audio and visual features. The FGSE module comprises a *Segment-wise Event Co-occurrence Modeling (SECM)* block and a *Local-Global Semantic Fusion (LGSF)* block. The SECM is designed to model inter-class dependencies among concurrent events within the same timestamp, enabling each event-specific feature to aggregate useful semantics from other relevant classes, where we also propose a novel *event co-occurrence loss function* to facilitate the learning of the event co-occurrence map. And the LGSF block considers event relations across temporal stamps by interacting the feature of each local segment with the global video feature, which is derived by averaging all the temporal features to provide more informative and noise-robust event semantics as events typically span consecutive temporal segments. Notably, the SECM and LGSF blocks can be applied to both intra-modality and cross-modality. More implementation details will be introduced in Sec. 3.2.

Our main contributions can be summarized as follows:

- We analyze the intractable semantic interference issue in the AVVP task and present a novel Multi-Modal Class-aware Semantic Enhancement (MM-CSE) network.
- We propose a class-aware feature decoupling module to decouple the semantically mixed audio/visual features into distinct event-specific and background features, facilitating event semantic perception from a more precise class level.
- We propose a fine-grained semantic enhancement module consisting of a segment-wise event co-occurrence block and a local-global semantic fusion block, further enhancing the event semantics of each segment. A new event co-occurrence loss function is introduced in this module.
- Extensive quantitative and qualitative experiments validate the effectiveness of our method, which achieves a new state-of-the-art audio-visual video parsing performance.

2 Related Work

Audio-Visual Learning aims to leverage both audio and visual modalities to enable machines to emulate the human perception process (Cheng et al. 2020; Mao et al. 2024; Shen et al. 2023; Li et al. 2024; Zhou et al. 2024b). Tasks such as sound source localization (Zhao et al. 2018; Qian et al. 2020), audio-visual segmentation (Zhou et al. 2022, 2024d; Guo et al. 2023) and event localization (Tian et al. 2018; Zhou et al. 2021; Zhou, Guo, and Wang 2022; Zhou et al. 2024a) explore audio-visual learning from spatial or temporal perspectives. Most of these prior studies presume synchronized audio and visual signals to be semantic-corresponding. However, this assumption does not hold in many situations, such as when the sound source is off-screen. In our studied AVVP task, the audio and visual signals may not always be temporally aligned. Our goal is to discern events within each modality, necessitating a more robust approach for articulating both intra-modal and inter-modal relations.

Audio-Visual Video Parsing aims to comprehensively identify and temporally localize events occurring in independent audio and visual modalities. In early research works (Tian, Li, and Xu 2020), this task is approached under a weakly supervised setting, where only the event label of the entire video is available for model training. To enhance supervision, some subsequent works have attempted to generate pseudo labels of events for independent audio and visual modalities at video-level (Wu and Yang 2021; Cheng et al. 2022) or more fine-grained segment-level (Zhou et al. 2023; Lai, Chen, and Yu-Chiang 2023; Fan et al. 2023; Zhou et al. 2024c). The

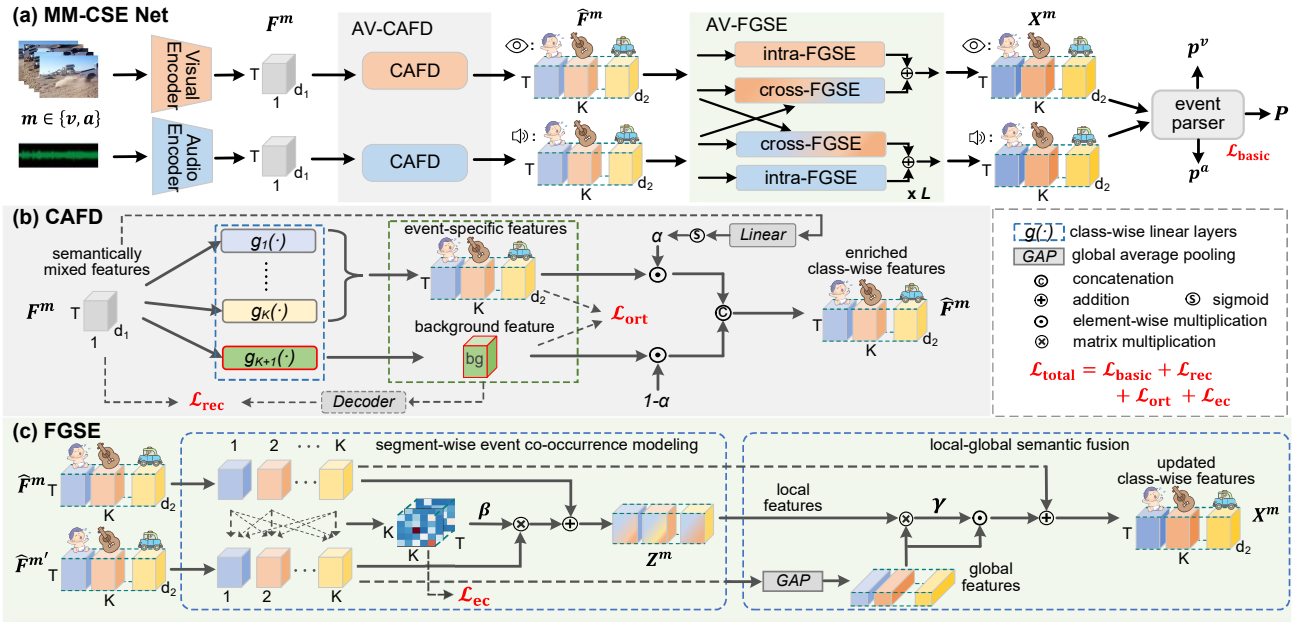

Figure 2: Framework Overview. (a) Our MM-CSE network primarily consists of two core modules: the audio-visual Class-Aware Feature Decoupling (AV-CAFD) and the Fine-Grained Semantic Enhancement (AV-FGSE). (b) The CAFD module decouples the encoded audio/visual features into distinct class-wise features, each representing a specific event or the background class. To ensure effective decoupling, we introduce a reconstruction loss $\mathcal{L}_{\text{rec}}$ and an orthogonality loss $\mathcal{L}_{\text{ort}}$. (c) The FGSE module further enhances the obtained class-wise features through two successive blocks: the Segment-wise Event Co-occurrence Modeling (SECM) block and the Local-Global Semantic Fusion (LGSF) block. The SECM encodes the relations among concurrent events within each timestamp, whereas the LGSF enhances the event semantics of local temporal segments by fusing relevant semantics of the global video. We also introduce an event co-occurrence loss $\mathcal{L}_{\text{ec}}$ to steer the learning of event co-occurrence in the SECM block. Notably, the FGSE module is applied to both intra-modality ('intra-FGSE') and cross-modality ('cross-FGSE').

generated segment-level pseudo labels can significantly improve the video parsing performance. Moreover, a variety of methods are proposed to encode more effective audio-visual representations for superior event parsing. The pioneer work HAN (Tian, Li, and Xu 2020) proposes a hybrid attention network that adopts the multi-head attention (Vaswani et al. 2017) mechanism to encode the intra-modal and cross-modal relations. Subsequently, some methods (Yu et al. 2022; Jiang et al. 2022; Zhang and Li 2023) are specially designed to better localize audio/visual events at varied temporal durations. To identify multiple events in audio or visual modalities, MGN (Mo and Tian 2022) utilizes multiple learnable class tokens to group the event semantics contained in each modality. Our method is significantly different from MGN as well as previous methods because we aim to address the semantic interference and our model relies on the decoupled class-wise features rather than the semantically mixed hidden features. More importantly, our approach uniquely accounts for a special *background* class, which is overlooked in all prior methods.

3 Our Approach

Given a video sequence comprising T non-overlapping pairs of audio and visual segments $\{a_t,\ v_t\}_{t=1}^{T}$, the AVVP task aims to predict all types of events within each segment pair, labeled as $(y_t^a, y_t^v, y_t^{av}) \in \mathbb{R}^{1\times K}$. Here, y_t^a, y_t^v and $y_t^{av} = y_t^a * y_t^v$ denote the audio, visual and audio-visual events, respectively, and K is the total number of event categories. Notably, each segment may contain multiple events or no events at all. During training, we can only access the manually annotated video-level label $\boldsymbol{Y} \in \mathbb{R}^{1\times K}$, depicting all events present in the video. However, the weak label $\boldsymbol{Y}$ does not specify the temporal segments or the modality in which these events occur. The most recent work (Lai, Chen, and Yu-Chiang 2023) provides high-quality pseudo labels for each modality at the segment-level, denoted as $\hat{\boldsymbol{y}}^a, \hat{\boldsymbol{y}}^v \in \mathbb{R}^{T\times K}$, offering fine-grained supervision. Fig. 2 illustrates the framework of our proposed MM-CSE. We provide its details in the next subsections.

3.1 Class-Aware Feature Decoupling

In the AVVP task, each audio/visual segment may contain multiple events; consequently, the event semantics are intermixed in the extracted audio/visual feature $\boldsymbol{F}^a/\boldsymbol{F}^v \in \mathbb{R}^{T\times d_1}$, resulting in semantic interference when performing intra-modal and cross-modal interactions, as discussed in Sec. 1.

To address this issue, we propose a CAFD module to explicitly decouple the semantically mixed features $\boldsymbol{F}^a, \boldsymbol{F}^v \in \mathbb{R}^{T\times d_1}$ into distinct class-aware features $\widetilde{\boldsymbol{F}}^a$, $\widetilde{\boldsymbol{F}}^v \in \mathbb{R}^{T\times(K+1)\times d_2}$, with '$K+1$' denoting the K *event* class plus an additional *background* class. We incorporate a background class for each segment to account for both the

event-related background context and segments that may only contain non-informative noise.

As shown in Fig. 2 (b), we accomplish the feature decoupling via K+1 independent linear layers $\{g_k\}_{k=1}^{K+1}$, obtaining the decoupled features $\widetilde{\boldsymbol{F}}^m = \{\widetilde{\boldsymbol{f}}_k^m\}_{k=1}^{K+1}$, where $m \in \{a, v\}$ denotes the audio and visual modalities, respectively. Here, we define the features $\{\widetilde{\boldsymbol{f}}_k^m\}_{k=1}^{K}$ as the event-specific features $\widetilde{\boldsymbol{F}}_{\text{e}}^m$ and $\widetilde{\boldsymbol{f}}_{K+1}^m$ as the background feature $\widetilde{\boldsymbol{F}}_{\text{bg}}^m$. It is worth noting that we introduce a reconstruction loss $\mathcal{L}_{\text{rec}}$ and an orthogonality loss $\mathcal{L}_{\text{ort}}$ on the decoupled features to guide effective feature decoupling. The former helps to maximize the semantic preservation of decoupled class-wise features compared to the original holistic feature, while the latter minimizes the relevance between the event-specific and the background features to make them more discriminative. We will elaborate on these loss functions in Sec. 3.3.

Considering some background information could also be helpful for event recognition, e.g., the background context may indicate this is an indoor event, we further dynamically integrate the background feature $\widetilde{\boldsymbol{F}}_{\text{bg}}^m$ into each event class feature $\widetilde{\boldsymbol{F}}_{\text{e}}^m$. Specifically, we first utilize linear layers and the Sigmoid function to generate a fusion weight vector $\alpha^m \in \mathbb{R}^{T\times 1}$ through the original holistic feature $\boldsymbol{F}^m$. In principle, α^m indicates the importance of the background in recognizing the current segments. Then, we employ a weighting mechanism to blend $\widetilde{\boldsymbol{F}}_{\text{bg}}^m$ with $\widetilde{\boldsymbol{F}}_{\text{e}}^m$, yielding the enriched class-wise feature $\widehat{\boldsymbol{F}}^m$ as follows,

$$\widehat{\boldsymbol{F}}^m = h[\alpha^m \cdot \widetilde{\boldsymbol{F}}_{\text{e}}^m; (1-\alpha^m)\cdot \widetilde{\boldsymbol{F}}_{\text{bg}}^m], \tag{1}$$

where $\widehat{\boldsymbol{F}}^m \in \mathbb{R}^{T\times K\times d_2}$, h is implemented by a linear layer followed by the ReLU function. Note that $\widetilde{\boldsymbol{F}}_{\text{bg}}^m$ is repeated K times to match the dimension of $\widetilde{\boldsymbol{F}}_{\text{e}}^m$ for the concatenation operation $[;]$ in Eq. 1, and we omit this for simplicity.

3.2 Fine-Grained Semantic Enhancement

After obtaining the enriched class-wise features $\widehat{\boldsymbol{F}}^m \in \mathbb{R}^{T\times K\times d_2}, m \in \{a, v\}$, we consider enhancing event semantics of each segment by exploring the semantic relevance from a more fine-grained class-level. To this end, we propose a FGSE module. Specifically, given the class-wise audio feature $\widehat{\boldsymbol{F}}^a$ and visual feature $\widehat{\boldsymbol{F}}^v$, our FGSE module encodes the intra-modal and cross-modal interactions as follows:

$$\boldsymbol{X}^m = \varphi_{\text{intra}}^m(\widehat{\boldsymbol{F}}^m, \widehat{\boldsymbol{F}}^m) + \varphi_{\text{cross}}^m(\widehat{\boldsymbol{F}}^m, \widehat{\boldsymbol{F}}^{\bar{m}}), \tag{2}$$

where $\boldsymbol{X}^m \in \mathbb{R}^{T\times K\times d_2}$ is the updated feature for the m modality, and $\bar{m} = \{a, v\}\backslash m$ denotes the counterpart modality. As illustrated in Fig. 2 (a) and (c), φ_{intra} and φ_{cross} share the same operations but have different parameters, each consisting of a *SECM* block and a *LGSF* block. We introduce the design principles and details of these two blocks next.

Segment-wise Event Co-occurrence Modeling (SECM). Each audio/visual segment may contain multiple events, indicating the event co-occurrence. For example, events such as 'people cheering' and 'people clapping' frequently co-occur within or across modalities, highlighting their relevant event semantics. Given that the class-wise features $\widehat{\boldsymbol{F}}^m \in \mathbb{R}^{T\times K\times d_2}$ are available, we design the SECM block to exploit the dependencies of such concurrent events. Specifically, for each segment, we measure the event co-occurrence by accessing the similarity (Vaswani et al. 2017) among K-class event features, computed as follows,

$$\boldsymbol{\beta}^{m,m'} = \text{Softmax}(\widehat{\boldsymbol{F}}^m(\widehat{\boldsymbol{F}}^{m'})^\top/\sqrt{d_2}), \tag{3}$$

where $m' \in \{m, \bar{m}\}$. The resulting $\boldsymbol{\beta}^{m,m'} \in \mathbb{R}^{T\times K\times K}$ indicates the event co-occurrence weight, where a larger value $\boldsymbol{\beta}_{:,i,j}^{m,m'}$ implies that the i-th event in modality m is more likely to coexist with the j-th event in modality m'. Then, each event-specific feature can be enhanced by aggregating relevant semantics from features of its concurrent events, computed by, $\boldsymbol{Z}^m = \widehat{\boldsymbol{F}}^m + \boldsymbol{\beta}^{m,m'}(\widehat{\boldsymbol{F}}^{m'})^\top$, where $\boldsymbol{Z}^m \in \mathbb{R}^{T\times K\times d_2}$ is the enhanced class-wise feature.

Local-Global Semantic Fusion (LGSF). The above SECM block performs inter-class co-occurrence modeling *within each local timestamp*. In order to harness relevant event semantics *across temporal timestamps*, we consider enhancing the semantics of each class within local segments by utilizing the global video information. We consider such a local-global interaction recognizing that the event semantics of the global video are typically more robust and informative, given that an event often spans consecutive temporal timestamps.

Specifically, we obtain the event feature of global video $\overline{\boldsymbol{G}}^{m'} \in \mathbb{R}^{1\times K\times d_2}$ via Global Average Pooling of $\widehat{\boldsymbol{F}}^{m'}$ along the temporal dimension. Then, we evaluate the semantic relevance between each local segment and the global video by computing their cosine similarity $\boldsymbol{\gamma}^{m,m'}$:

$$\boldsymbol{\gamma}^{m,m'} = \langle \|\boldsymbol{Z}^m\|, \|\overline{\boldsymbol{G}}^{m'}\| \rangle, \tag{4}$$

where $\|\cdot\|$ denotes the L2-normalization, $\boldsymbol{\gamma}^{m,m'} \in \mathbb{R}^{T\times K\times 1}$. The element $\boldsymbol{\gamma}_{:,k}^{m,m'} \in \mathbb{R}^{T\times 1}$ indicates the feature similarity between T local segments with the global video for the k-th event class. Afterward, the feature of each local segment can be enhanced by aggregating relevant event semantics from global video feature via $\boldsymbol{\gamma}^{m,m'}$, calculated as, $\boldsymbol{X}^m = \widehat{\boldsymbol{F}}^m + \boldsymbol{\gamma}^{m,m'} \odot \overline{\boldsymbol{G}}^{m'}$.

For convenience, we abbreviate the above operations of the SECM and the LGSF blocks as a single FGSE layer:

$$\boldsymbol{X}^a, \boldsymbol{X}^v = \text{FGSE}(\widehat{\boldsymbol{F}}^a, \widehat{\boldsymbol{F}}^v), \tag{5}$$

Then, we utilize L stacked FGSE layers to iteratively enhance the class-wise audio and visual features:

$$\boldsymbol{X}_l^a, \boldsymbol{X}_l^v = \text{FGSE}(\boldsymbol{X}_{l-1}^a, \boldsymbol{X}_{l-1}^v), \tag{6}$$

where $l = \{1, 2, ..., L\}$ is the layer index and $\boldsymbol{X}_0^m = \widehat{\boldsymbol{F}}^m$.

3.3 Loss Functions

The audio and visual features obtained by the final FGSE layer, $\boldsymbol{X}_L^m \in \mathbb{R}^{T\times K\times d_2}(m \in \{a, v\})$, are fed into the event parser to predict the segment-level event probabilities $\boldsymbol{p}^m \in \mathbb{R}^{T\times K}$. The video-level prediction for the entire video $\boldsymbol{P} \in \mathbb{R}^{1\times K}$ can be obtained from $\boldsymbol{p}^a$ and $\boldsymbol{p}^v$ through an attentive

multi-modal multi-instance (MMIL) pooling (Tian, Li, and Xu 2020) mechanism.

Basic classification loss $\mathcal{L}_{\text{basic}}$. Given the segment-level and video-level predictions, the basic loss function $\mathcal{L}_{\text{basic}}$ utilizes the binary cross entropy (BCE) loss to align them with the ground truths, computed as,

$$\mathcal{L}_{\text{basic}} = \text{BCE}(\boldsymbol{p}^a, \hat{\boldsymbol{y}}^a) + \text{BCE}(\boldsymbol{p}^v, \hat{\boldsymbol{y}}^v) + \text{BCE}(\boldsymbol{P}, \boldsymbol{Y}), \quad (7)$$

Additionally, our optimization objective involves two loss functions aimed at facilitating effective class-aware feature decoupling ($\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{ort}}$) and a novel loss function $\mathcal{L}_{\text{ec}}$ designed to regularize the event co-occurrence modeling in the SECM module. Next, we elaborate on each of them.

Reconstruction loss $\mathcal{L}_{\text{rec}}$. To maximize the preservation of semantic information within the decoupled features, we feed the decoupled class-wise features $\widetilde{\boldsymbol{F}}^m$ into a *Decoder* to reconstruct the original holistic feature $\boldsymbol{F}^m$. The *Decoder* is implemented by two linear layers inserted by a ReLU activation. Then, we calculate the reconstruction loss $\mathcal{L}_{\text{rec}}$ by the mean squared error (MSE):

$$\mathcal{L}_{\text{rec}} = \sum_m \text{MSE}(\textit{Decoder}(\widetilde{\boldsymbol{F}}^m), \boldsymbol{F}^m), \quad (8)$$

Orthogonality loss $\mathcal{L}_{\text{ort}}$. In addition to $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{ort}}$ is introduced to promote dissimilarity between the background feature $\widetilde{\boldsymbol{F}}^m_{\text{bg}} \in \mathbb{R}^{T\times 1\times d_2}$ and the event-specific feature $\widetilde{\boldsymbol{F}}^m_{\text{e}} \in \mathbb{R}^{T\times K\times d_2}$, enhancing the isolation of the background information for improved event prediction. $\mathcal{L}_{\text{ort}}$ is calculated as follows,

$$\mathcal{L}_{\text{ort}} = \frac{1}{TK}\sum_m \sum_{t=1}^{T}\sum_{k=1}^{K} \langle \|\widetilde{\boldsymbol{F}}^m_{\text{bg}}\|, \|\widetilde{\boldsymbol{F}}^m_{\text{e}}\| \rangle, \quad (9)$$

where $\langle \|\widetilde{\boldsymbol{F}}^m_{\text{bg}}\|, \|\widetilde{\boldsymbol{F}}^m_{\text{e}}\| \rangle$ is the cosine similarity matrix which has the dimension of $T \times K$.

Event co-occurrence loss $\mathcal{L}_{\text{ec}}$. Within the SECM module, the $\boldsymbol{\beta}^{m,m'} \in \mathbb{R}^{T\times K\times K}$ reflects the learned event co-occurrence relations between modality m and modality m'. We can also obtain the event co-occurrence matrix $\boldsymbol{M}^{m,m'} \in \mathbb{R}^{T\times K\times K}$ from the segment-level labels $\hat{\boldsymbol{y}}^m, \hat{\boldsymbol{y}}^{m'} \in \mathbb{R}^{T\times K}$ as ground truth, i.e.,

$$\boldsymbol{M}^{m,m'}_{t,(i,j)} = \begin{cases} 1, & \text{if } \hat{\boldsymbol{y}}^m_{t,i} = \hat{\boldsymbol{y}}^{m'}_{t,j} = 1 \\ 0, & other \end{cases} \quad (10)$$

Then, the $\mathcal{L}_{\text{ec}}$ can be computed by,

$$\mathcal{L}_{\text{ec}} = \text{MSE}(\boldsymbol{\beta}^{m,m'}, \boldsymbol{M}^{m,m'}), \quad (11)$$

In summary, the overall objective $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{basic}} + \mathcal{L}_{\text{rec}} + \lambda_1\mathcal{L}_{\text{ort}} + \lambda_2\mathcal{L}_{\text{ec}}. \quad (12)$$

where λ_1 and λ_2 are the balancing hyperparameters.

4 Experiments

4.1 Experimental Setups

Dataset & Metrics. Following prior works (Jiang et al. 2022; Lai, Chen, and Yu-Chiang 2023), our experiments are conducted on the *Look, Listen, and Parse* (LLP) (Tian, Li, and Xu 2020) dataset, which is currently the sole standard dataset used for the AVVP task. It comprises 11,849 videos covering 25 common categories. Following the official data splits, the dataset is divided into 10,000 videos for training, 649 for validation, and 1,200 for testing. For the evaluation metrics, we adopt F-scores across all types of events: audio events (A), visual events (V), and audio-visual events (AV), at both segment-level and event-level. Furthermore, 'Type' is the averaged F-scores of A, V, and AV metrics. 'Event' considers audio and visual events simultaneously for each sample.

Implementation details. *1) Feature extraction.* Following the previous SOTA method VALOR, video frames are sampled at 8 FPS and the pretrained CLIP (Radford et al. 2021) and R(2+1)D (Tran et al. 2018) models are utilized to extract the 2D and 3D features, respectively, which are then concatenated to form the visual features. The corresponding audio features are extracted using the pretrained CLAP (Wu et al. 2023) model. *2) Training configurations.* Our model is trained for 60 epochs with a batch size of 64 using AdamW optimizer, with an initial learning rate of 3e-4 and a weight decay of 1e-3. Feature dimensions d_1 and d_2 are set to 256 and 128, respectively. We use $L = 4$ stacked FGSE layers. The hyperparameters λ_1 and λ_2 in Eq. 12 are empirically set to 0.1. The code will be publicly available in https://github.com/PengchengZhao1001/MM-CSE.

4.2 Comparison with State-of-the-Arts

We quantitatively compare our proposed MM-CSE method with prior works. Notably, previous works typically utilize VGGish (Hershey et al. 2017), pretrained on AudioSet (Gemmeke et al. 2017) to extract audio features and employ the pretrained ResNet-152 (He et al. 2016) and R(2+1)D (Tran et al. 2018) models to extract visual features. For a fair comparison, we employ the same audio and visual backbones for feature extraction, reporting the performances in the upper part of Table 1. Our method sets a new benchmark with a remarkable 61.5% average parsing performance, surpassing the prior SOTA VALOR (Lai, Chen, and Yu-Chiang 2023), especially in audio event parsing with a notable $\sim$2% gain in both segment-level and event-level metrics. Furthermore, we compare our method against VALOR using the same CLAP and CLIP features. As presented in the final two rows of Table 1, our method continues to outperform VALOR in all types of event parsing, surpassing it by 2.2% on the average performance. These results demonstrate the superiority of our proposed method. In the subsequent section, we validate the effectiveness and advantages of each core component.

4.3 Ablation Studies

In this section, we conduct ablation studies to validate the key designs in our model. *Notably, due to space limitations, we present the average result of segment-level and event-level metrics for the experiments.* More detailed results of each metric are further provided in our Supplementary Material.

Ablations on the CAFD module. We validate this module by exploring various feature decoupling strategies, as illustrated in Table 2: #1) We implement our model directly using the semantically mixed features without decoupling; #2) We decouple the features into only K event-specific classes

Methods	Segment-level					Event-level					Avg.
	A	V	AV	Type	Event	A	V	AV	Type	Event	
HAN (Tian, Li, and Xu 2020)	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0	51.0
CVCMS (Lin et al. 2021)	59.2	59.9	53.4	57.5	58.1	51.3	55.5	46.2	51.0	49.7	54.2
MA (Wu and Yang 2021)	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6	55.5
MGN (Mo and Tian 2022)	60.2	61.9	55.5	59.2	58.7	50.9	59.7	49.6	53.4	49.9	55.9
MM-Pyr (Yu et al. 2022)	61.1	60.3	55.8	59.7	59.1	53.8	56.7	49.4	54.1	51.2	56.1
JoMoLD (Cheng et al. 2022)	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5	57.3
DHHN (Jiang et al. 2022)	61.4	63.4	56.8	60.5	59.5	54.6	60.8	51.1	55.5	53.3	57.7
CMPAE (Gao, Chen, and Xu 2023)	64.2	66.4	59.2	63.3	62.8	56.6	63.7	51.8	57.4	55.7	60.1
LSLD (Fan et al. 2023)	62.7	67.1	59.4	63.1	62.2	55.7	64.3	52.6	57.6	55.2	60.0
VALOR (Lai, Chen, and Yu-Chiang 2023)	62.8	66.7	60.0	63.2	62.3	57.1	63.9	54.4	58.5	55.9	60.5
MM-CSE (ours)	**65.0**	**66.8**	**60.0**	**63.9**	**64.2**	**59.1**	**64.1**	**54.9**	**59.4**	**57.6**	**61.5**
VALOR* (Lai, Chen, and Yu-Chiang 2023)	68.1	68.4	61.9	66.2	66.8	61.2	64.7	55.5	60.4	59.0	63.2
MM-CSE* (ours)	**69.5 (+1.4)**	**71.3 (+2.9)**	**64.2 (+2.3)**	**68.3 (+2.1)**	**68.9 (+2.1)**	**63.0 (+1.8)**	**67.5 (+2.8)**	**57.8 (+2.3)**	**62.7 (+2.3)**	**61.1 (+2.1)**	**65.4 (+2.2)**

Table 1: Comparison with state-of-the-art methods. The results shown in the upper Table are obtained by using audio and visual features respectively extracted by VGGish and ResNet models, while results in the last two rows ($*$) denote that the CLAP and CLIP features are used. 'Avg.' is the average result of all ten metrics.

#	EVE.	BG	A	V	AV	Type	Event	Avg.
1	×	×	64.4	66.2	58.6	63.1	62.5	63.0
2	✓	×	65.2	68.1	59.8	64.3	63.7	64.2
3	✓	✓	**66.3**	**69.4**	**61.0**	**65.5**	**65.0**	**65.4**

Table 2: Ablation study on the CAFD module. 'EVE.' denotes K event-specific classes, while 'BG' stands for an additional background class.

Strategy		A	V	AV	Type	Event	Avg.
full		**66.3**	**69.4**	**61.0**	**65.5**	**65.0**	**65.4**
blocks	w/o SECM	62.7	65.3	57.1	61.7	61.5	61.6
	w/o LGSF	64.5	66.8	59.0	63.4	62.6	63.2
modes	w/o intra-FGSE	64.6	68.5	59.6	64.2	63.5	64.1
	w/o cross-FGSE	59.7	67.2	54.3	60.4	61.6	60.6

Table 3: Ablation study on the FGSE module. We explore the impacts of each block and each mode.

('EVE.'). #3) We decouple the features into both K event-specific classes and an extra background class ('EVE.'+'BG'). Our experiments reveal that the model exhibits the poorest performance when removing the CAFD module (#1). However, feature decoupling for event classes leads to a notable enhancement by an average of 1.2% (#2). Additionally, optimal performance is achieved by further including the background class (#3). These findings suggest that it is crucial to consider feature decoupling for both the event-specific and background classes. The decoupled features ensure more nuanced semantic modeling via class-level interactions.

Ablations on the FGSE module. As introduced in Sec. 3.2, the proposed FGSE module comprises two blocks: the segment-wise co-occurrence modeling (SECM) and the local-global semantic fusion (LGSF). Besides, the FGSE operates in two modes: intra-modality and cross-modality. We perform ablations on each block and each mode to assess their impacts. As shown in Table 3, the model's performance experiences a notable decrease when either block or mode is removed. This demonstrates the effectiveness of each component of the FGSE module. From the last two rows of the Table, we can also conclude that considering the cross-modal interactions is more important for the AVVP task.

#	$\mathcal{L}_{\text{basic}}$	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{ort}}$	$\mathcal{L}_{\text{ec}}$	A	V	AV	Type	Event	Avg.
1	✓	×	×	×	65.6	67.8	59.8	64.4	64.1	64.3
2	✓	✓	×	×	65.5	68.1	59.4	64.3	64.1	64.3
3	✓	×	✓	×	65.3	68.4	59.6	64.4	64.3	64.4
4	✓	✓	✓	×	65.8	69.0	60.0	64.9	64.7	64.8
5	✓	✓	✓	✓	**66.3**	**69.4**	**61.0**	**65.5**	**65.0**	**65.4**

Table 4: Ablation study on the loss functions.

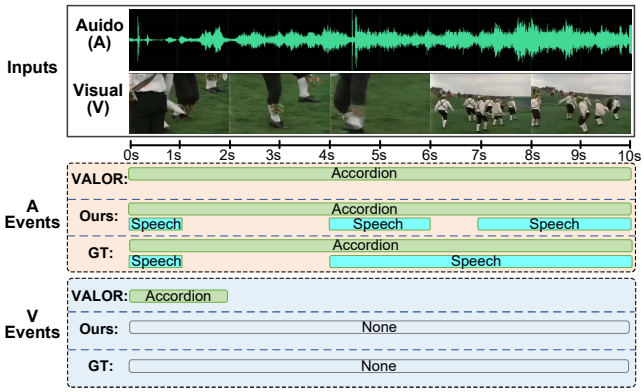

Figure 3: Qualitative comparisons. Compared to the previous SOTA method VALOR, our method performs better in identifying multiple overlapping events and recognizing or utilizing the background information.

Ablations on the loss functions. We explore the impact of each loss by ablating them to train the model. As shown in Table 4, the model trained solely with the basic $\mathcal{L}_{\text{basic}}$ already achieves satisfactory performance, with an average metric of 64.3% (#1). Notably, our model trained in this setting al-

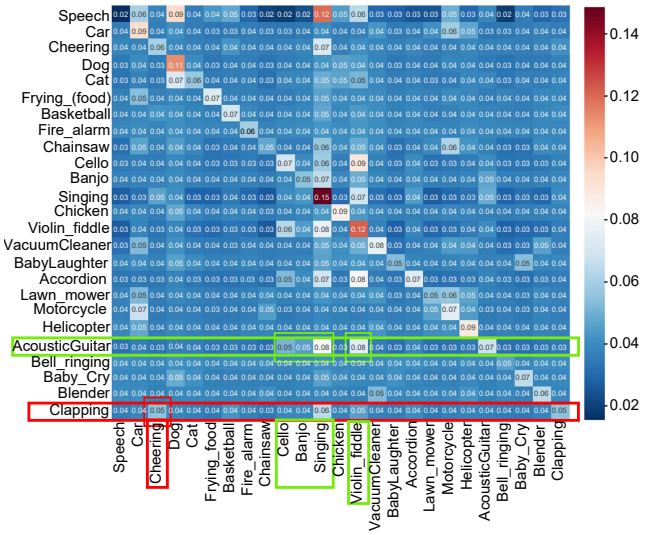

Figure 4: Visualization example of the learned event co-occurrence map.

ready surpasses the previous SOTA method VALOR (63.2%), demonstrating the superiority of our network design. Moreover, it is observed that using $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{ort}}$ independently does not yield obvious benefits (#2 and #3). However, combining these two losses enhances the performance, suggesting their complementarity for effective class-aware feature decoupling (#4). Finally, the introduction of $\mathcal{L}_{\text{ec}}$ further improves the performance by 0.6% as it provides additional supervision during event co-occurrence modeling (#5).

4.4 Qualitative Results

Qualitative example of audio-visual video parsing. We first present a parsing example to intuitively compare our proposed method with the previous SOTA method VALOR. As shown in Fig. 3, there are two audio events: *Accordion* and *Speech*. VALOR successfully identifies *Accordion* event but completely misses the *Speech* event. Conversely, our method satisfactorily localizes segments containing both events. An interesting observation is evident in the visual track where VALOR incorrectly recognizes that the event *Accordion* occurs in the first two segments. However, the visual frames actually depict people's feet on the playground. This misunderstanding may arise from semantic interference from the audio events related to the *Accordion*, as VALOR performs cross-modal interaction directly relying on semantically mixed audio and visual features. In contrast, our method accurately classifies the first two segments as backgrounds. Our method utilizes decoupled class-aware features for intra- and cross-modal interactions, enabling each class's feature to aggregate only similar semantics from relevant classes. Additionally, introducing an additional background class during the feature decoupling process helps the model distinguish the background from diverse events, preventing the model from consistently falling into predefined event categories.

Visualization of learned event co-occurrence map. We then present the learned event co-occurrence map, derived by averaging the $\boldsymbol{\beta}^{av} \in \mathbb{R}^{T\times K\times K}$ across all test videos along the temporal dimension. In Fig. 4, the visualization illustrates that our model effectively captures the co-occurrence among relevant events. For instance, as highlighted by the green boxes, the audio event *Acoustic_guitar* exhibits higher correspondence with the visual event *Singing* and some instrumental events. Similarly, the audio event *Clapping* correlates well with the visual event *Cheering* (red boxes). These observations align with common scenarios where such audio-visual event pairs frequently co-occur. Modeling such event co-occurrence enhances the model's capability to perceive concurrent events.

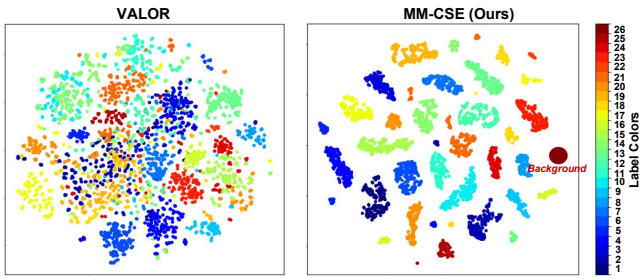

Figure 5: Visualization of our decoupled class-wise features. Each color represents one class.

Visualization of the decoupled class-wise features. Finally, we visualize the distributions of our decoupled features using the t-SNE (Van der Maaten and Hinton 2008). Here, we take the audio modality for demonstration. In Fig. 5, we also compare our method with the previous SOTA method VALOR which exhibits semantically mixed audio representations across different event classes. In contrast, our decoupled features, including both event-specific and background features, demonstrate pronounced intra-class compactness and inter-class separation. These results validate that our method effectively decouples the semantically mixed hidden features into separate class-aware features. Such class-wise features allow more accurate interactions from class-level for both intra- and cross-modality, thereby enhancing multi-class event classification in the AVVP task.

5 Conclusion

For the audio-visual video parsing task, existing approaches directly utilize semantically mixed holistic audio and visual features for modeling intra- and cross-modal relations, which could result in semantic interference. In this paper, we propose decoupling the semantically mixed features into distinct event-specific and background-specific class-wise features. These decoupled features allow for precise event semantic learning for each segment by aggregating positive supports from features of highly relevant classes. Specifically, the proposed FGSE module first encodes the inter-class dependencies among concurrent events within each timestamp, and then enriches each local segment with matched global video semantics. Additionally, several loss functions are introduced for effective feature decoupling and event co-occurrence modeling. Extensive experiments verify the superiority of our class-aware semantic enhancement method.

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers for their invaluable comments and insightful suggestions. This work was supported by the National Natural Science Foundation of China (61972127, 62272142, 62272144, 72188101, and 62020106007), the Major Project of Anhui Province (2408085J040, 202203a05020011), and the Fundamental Research Funds for the Central Universities (JZ2024HGTG0309, JZ2024AHST0337, and JZ2023YQTD0072).

References

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 431–448.

Cheng, Y.; Wang, R.; Pan, Z.; Feng, R.; and Zhang, Y. 2020. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 3884–3892.

Fan, Y.; Wu, Y.; Lin, Y.; and Du, B. 2023. Revisit Weakly-Supervised Audio-Visual Video Parsing from the Language Perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1–13.

Gao, J.; Chen, M.; and Xu, C. 2023. Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18827–18836.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.

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*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 131–135.

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 *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 719–727.

Lai, Y.-H.; Chen, Y.-C.; and Yu-Chiang, F. W. 2023. Modality-Independent Teachers Meet Weakly-Supervised Audio-Visual Event Parser. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1–19.

Li, Z.; Guo, D.; Zhou, J.; Zhang, J.; and Wang, M. 2024. Object-Aware Adaptive-Positivity Learning for Audio-Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 3306–3314.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 1–13.

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 *Proceedings of the ACM international conference on multimedia (ACM MM)*, 6607–6616.

Mo, S.; and Tian, Y. 2022. Multi-modal Grouping Network for Weakly-Supervised Audio-Visual Video Parsing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1–12.

Qian, R.; Hu, D.; Dinkel, H.; Wu, M.; Xu, N.; and Lin, W. 2020. Multiple sound sources localization from coarse to fine. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 292–308.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10585–10596.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 436–454.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 247–263.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6450–6459.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 1–11.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio

pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Wu, Y.; and Yang, Y. 2021. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1326–1335.

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 *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 6241–6249.

Zhang, J.; and Li, W. 2023. Multi-Modal and Multi-Scale Temporal Fusion Architecture Search for Audio-Visual Video Parsing. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 3328–3336.

Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; and Torralba, A. 2018. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 570–586.

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: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 *European Conference on Computer Vision (ECCV)*, 1–22.

Zhou, J.; Guo, D.; and Wang, M. 2022. Contrastive Positive Sample Propagation along the Audio-Visual Event Line. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1–18.

Zhou, J.; Guo, D.; Zhong, Y.; and Wang, M. 2023. Improving audio-visual video parsing with pseudo visual labels. 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. *International Journal of Computer Vision (IJCV)*, 1–22.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2024d. Audio-Visual Segmentation with Semantics. *International Journal of Computer Vision (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 *Proceedings of the European Conference on Computer Vision (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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8436–8444.