MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

UWAV: Uncertainty-weighted Weakly-supervised Audio-Visual Video Parsing

Lai, Yung-Hsuan; Ebbers, Janek; Wang, Yu-Chiang Frank; Germain, François G; Jones, Michael J.; Chatterjee, Moitreya

TR2025-072 June 07, 2025

Abstract

Audio-Visual Video Parsing (AVVP) entails the challenging task of localizing both uni-modal events (i.e., those occurring exclusively in either the visual or acoustic modality of a video) and multi-modal events (i.e., those occur- ring in both modalities concurrently). Moreover, the prohibitive cost of annotating training data with the class labels of all these events, along with their start and end times, imposes constraints on the scalability of AVVP techniques unless they can be trained in a weakly-supervised setting, where only modality-agnostic, video-level labels are available in the training data. To this end, recently pro- posed approaches seek to generate segment-level pseudo- labels to better guide model training. However, the absence of inter-segment dependencies when generating these pseudo-labels and the general bias towards predicting labels that are absent in a segment limit their performance. This work proposes a novel approach towards overcoming these weaknesses called Uncertainty-weighted Weakly-supervised Audio-visual Video Parsing (UWAV). Addition- ally, our innovative approach factors in the uncertainty associated with these estimated pseudo-labels and incorporates a feature mixup based training regularization for improved training. Empirical results show that UWAV outperforms state-of-the-art methods for the AVVP task on multiple metrics, across two different datasets, attesting to its effectiveness and generalizability.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

UWAV: Uncertainty-weighted Weakly-supervised Audio-Visual Video Parsing

Yung-Hsuan Lai[1,†], Janek Ebbers[3], Yu-Chiang Frank Wang[1,2], François Germain[3],
Michael Jeffrey Jones[3], Moitreya Chatterjee[3,‡],
[1] Graduate Institute of Communication Engineering, National Taiwan University [2] NVIDIA, Taiwan
[3] Mitsubishi Electric Research Labs (MERL)
[†]r10942097@ntu.edu.tw [‡]chatterjee@merl.com

Abstract

*Audio-Visual Video Parsing (AVVP) entails the challenging task of localizing both uni-modal events (*i.e.*, those occurring exclusively in either the visual or acoustic modality of a video) and multi-modal events (*i.e.*, those occurring in both modalities concurrently). Moreover, the prohibitive cost of annotating training data with the class labels of all these events, along with their start and end times, imposes constraints on the scalability of AVVP techniques unless they can be trained in a weakly-supervised setting, where only modality-agnostic, video-level labels are available in the training data. To this end, recently proposed approaches seek to generate segment-level pseudo-labels to better guide model training. However, the absence of inter-segment dependencies when generating these pseudo-labels and the general bias towards predicting labels that are absent in a segment limit their performance. This work proposes a novel approach towards overcoming these weaknesses called Uncertainty-weighted Weakly-supervised Audio-visual Video Parsing (UWAV). Additionally, our innovative approach factors in the uncertainty associated with these estimated pseudo-labels and incorporates a feature mixup based training regularization for improved training. Empirical results show that UWAV outperforms state-of-the-art methods for the AVVP task on multiple metrics, across two different datasets, attesting to its effectiveness and generalizability.*

1. Introduction

Events that occur in the real world, often leave their imprint on the acoustic and visual modalities. Humans rely heavily on the synergy between their senses of sight and hearing to interpret such events. Audio-visual learning, which seeks to equip machines with a similar synergy, has emerged as one of the most important research areas within the multimodal machine learning community. It aims to leverage both these senses (modalities) jointly, to enhance machine perception and understanding of real-world events. Various audio-visual learning tasks have been studied towards this end, including audio-visual segmentation [19, 49], sound source localization [24, 32], audio-visual event localization [29, 34], and audio-visual sound separation [3, 7, 43]. However, many of these tasks assume that audio and visual streams would always be temporally aligned. This assumption often fails in real-world scenarios, where the sonic and visual imprints of events may not perfectly overlap. For instance, one might hear an emergency siren approaching from a distance before it appears in the field of view.

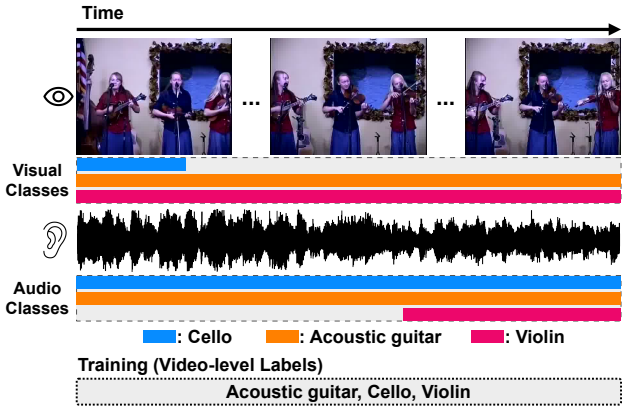

Figure 1. **A weakly-supervised AVVP task example.** Events, considered in this task, might be unimodal or multimodal. Even multimodal events, may not be temporally aligned in the audio and visual modalities, *e.g.* the cello might only be visible in the first few seconds but might produce music, throughout the video.

In this work, in order to better understand the events occurring in a video, we explore the task of Audio-Visual Video Parsing (AVVP) [35]. Its goal is to recognize and localize all audio, visual, and audio-visual events occurring in the video. See Figure 1 for an example of this task. The task setup is to perform this prediction for every one-second segment of a video. This task poses two principal challenges, from a machine learning standpoint: (i) The audio and visual events, that occur, might not be temporally aligned,

e.g. if an event becomes audible before its source enters the camera field of view, or the sound source is not visible at all, and (ii) due to the high costs of annotating video segments with per-segment labels, only modality-agnostic, video-level labels are provided during training, *i.e*., these labels specify which events occur in a video but lack details about the segments or the modality in which they occur.

Prior works in the area can be grouped into two orthogonal research directions. The first focuses on enhancing model architectures [23, 45, 51]. Despite advancements in this direction, the absence of fine-grained labels to guide the model during training continues to pose an impediment towards the generalizability of such models. As a result, recent approaches have focused on the second direction of research which aims at generating richer pseudo-labels for improved training, either at the video-level [8, 39] or segment-level [9, 17, 26, 50]. In particular, Rachavarapu et al. [26] propose prototype-based pseudo-labeling (PPL), which seeks to train a pseudo-label generation module in conjunction with a core inference module. However, due to the lack of sufficient training data, this method struggles to generalize. On the other hand, VPLAN [50], VALOR [17], and LSLD [9] leverage large-scale pre-trained foundation models, such as CLIP [28] and CLAP [40], along with ground-truth video-level labels to generate segment-level pseudo-labels for each of the two modalities. Audio/Visual segments (*e.g*. the audio corresponding to the segment in question and the visual frame at the center of the segment) are fed into CLAP/CLIP, one segment at a time, to generate these pseudo-labels. Despite the significant improvement that these pseudo-label generation methods achieved, the correctness of the generated labels is still limited, constrained primarily by the lack of understanding of inter-segment dynamics. For instance, if a crowd is *cheering* in a segment of the video, it is more likely that the crowd might also be *clapping* right before or after.

To address the oversight of inter-segment dependencies and other shortcomings in existing pseudo-label generation methods, we introduce a novel, uncertainty-based, weakly-supervised, video parsing model called UWAV (Uncertainty-weighted Weakly-supervised Audio-visual Video Parsing), capable of generating improved segment-level pseudo-labels for better training of the inference module. We resort to transformer modules [38] to equip UWAV with the ability to capture temporal relationships between segments and pre-train it on a large-scale, supervised audio-visual event localization dataset [12]. Subsequently, this pre-trained model is used to generate segment-level pseudo-labels for each modality on a target, small-scale dataset which only provides weak (*i.e*., video-level) supervision. Such a design permits a more holistic understanding of the video, resulting in more accurate pseudo-labels. Additionally, UWAV factors in the uncertainty associated with these estimated pseudo-labels in its optimization. That uncertainty is the result of the shift in the domain of the target dataset, insufficient model capacity, *etc*. and is computed at training time as the confidence scores associated with these labels. To further enhance the model's ability to learn in this small-scale, weakly-supervised data regime, we also employ a feature mixup strategy. This approach adds more regularization constraints by training on mixed segment features alongside interpolated pseudo-labels, which not only lessens the influence of noise but also enriches the training data, thereby reducing overfitting. Moreover, UWAV addresses a critical class imbalance issue in the pseudo-label enriched training data, viz. most event classes in any given segment of a video are absent/negative (*i.e*., they do not occur), while only a handful of them do. This creates a natural bias in the training set, making it difficult to learn the positive events. To counter this, we propose a class-frequency aware re-weighting strategy that lays greater emphasis on the accurate classification and localization of positive events. By incorporating these components into its design, our proposed method (UWAV) outperforms competing state-of-the-art approaches across two different datasets, viz. Look, Listen, and Parse (LLP) [35] and the Audio-Visual Event Localization (AVE) [34], on multiple evaluation metrics.

In summary, our contributions are the following:

- We introduce a novel, weakly-supervised method called UWAV, capable of synthesizing temporally coherent pseudo-labels for the AVVP task.
- To the best of our knowledge, ours is the first method for the AVVP task, which factors in the uncertainty associated with the estimated pseudo-labels while also regularizing it with a feature mixup strategy.
- UWAV outperforms competing state-of-the-art approaches for the AVVP task, across two different datasets on multiple metrics which attest to its generalizability.

2. Related Works

Audio-Visual Learning: Audio-visual learning has emerged as an area of active research, aiming to develop models that synergistically integrate information from both audio and visual modalities for improved perception and understanding. Towards this end, various audio-visual tasks have been explored by the community so far, such as audio-visual segmentation [19, 22, 42, 49], sound source localization [15, 16, 24, 32], event localization [29, 34, 48], navigation [4–6, 21, 44, 46], generation [25, 30, 41], question answering [11, 18, 33, 47], and sound source separation [2, 7, 36, 43]. In this work, we focus on the task of audio-visual video parsing (AVVP) where the goal is to temporally localize events occurring in a video. Unlike many other audio-visual learning tasks, AVVP does not assume that events are always aligned across

modalities. Some events could be exclusively uni-modal while others may have an audio-visual signature, which requires complex reasoning.

Audio-Visual Video Parsing (AVVP): To address the challenges of the AVVP task [17, 26, 35], Tian et al. [35] proposed a Hybrid Attention Network (HAN) and a learnable Multi-modal Multiple Instance Learning (MMIL) pooling module. The HAN model facilitates the exchange of information within and across modalities using self-attention and cross-attention layers, while the MMIL pooling module aggregates segment-level event probabilities from both modalities to produce video-level probabilities. Building on this foundation, recent works advanced the field from the following two perspectives. The first group of studies [23, 45, 51] focuses on enhancing model architectures. In particular, Mo and Tian [23] proposed the Multi-modal Grouping Network (MGN) to explicitly group semantically similar features within each modality to improve the reasoning process, while Yu et al. [45] proposed the Multi-modal Pyramid Attentional Network (MM-Pyramid) to capture events of varying durations by extracting features at multiple temporal scales. Our proposed method is orthogonal to this line of research and can be integrated with any of these backbones.

The second direction focuses on generating pseudo-labels for improved training, either at the video-level [8, 39] or the segment-level [9, 17, 26, 50]. VPLAN [50], VALOR [17], and LSLD [9] utilize pre-trained CLIP [28] and CLAP [40] along with ground-truth video-level labels to predict pseudo-labels for each modality on a per-segment basis. In contrast, PPL [26] uses the HAN model itself to generate pseudo-labels by constructing prototype features for each class and assigning labels to each segment based on the similarity between its feature and the prototype features. While these pseudo-label generation methods have substantially improved model performance on the AVVP task, they still exhibit some limitations. For instance, to derive accurate pseudo-labels, PPL might require a large enough training set to learn good prototype features, which might pose challenges when applied to smaller datasets. Our proposed method overcomes this problem. On the other hand, methods that leverage CLIP and CLAP to generate pseudo-labels often ignore temporal relationships between segments or the uncertainty associated with these labels. Our work also seeks to plug this void.

3. Preliminaries

Problem Formulation: The AVVP task [35] aims to localize all visible and/or audible events in each one-second segment of a video. Specifically, an audible video is split into T one-second segments, denoted as $\{V_t, A_t\}_{t=1}^{T}$. Each segment is annotated with a pair of ground-truth labels $y_t^v \in \{0,1\}^C, y_t^a \in \{0,1\}^C$, where y_t^v denotes visual events, y_t^a denotes audio events, and C denotes the total number of events in the pre-defined event set of the data. However, owing to the weakly-supervised nature of the task setup (y_t^v, y_t^a) are unavailable during training. Instead, only the modality-agnostic, video-level labels $y \in \{0,1\}^C$ are available, where 1 indicates the presence of an event at any time (either in the audio or visual stream or both) while 0 indicates an event's absence in the video.

Pseudo-Label Based AVVP Framework: The Hybrid Attention Network (HAN) [35] is a commonly used model for the AVVP task. The model works by first utilizing pre-trained visual and audio backbones to extract features from the visual and audio segments respectively, which are then projected to two d-dimensional feature spaces. The resulting visual segment-level features are denoted by $F^v = \{f_t^v\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$, while the audio segment-level features are denoted by $F^a = \{f_t^a\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$. These features are provided as input to the HAN model. In the model, information across segments within a modality and across modalities is exchanged through self-attention and cross-attention layers, as shown below:

$$\tilde{f}_t^v = f_t^v + \underbrace{\text{Attn}(f_t^v, F^v, F^v)}_{\text{Self-Attention}} + \underbrace{\text{Attn}(f_t^v, F^a, F^a)}_{\text{Cross-Attention}}, \quad (1)$$

$$\tilde{f}_t^a = f_t^a + \underbrace{\text{Attn}(f_t^a, F^a, F^a)}_{\text{Self-Attention}} + \underbrace{\text{Attn}(f_t^a, F^v, F^v)}_{\text{Cross-Attention}}, \quad (2)$$

where $\text{Attn}(Q, K, V)$ denotes the standard multi-head attention mechanism [38]. Finally a classifier, shared across both modalities, transforms the visual segment-level features $\tilde{F}^v = \{\tilde{f}_t^v\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$ (*resp.* audio segment-level features $\tilde{F}^a = \{\tilde{f}_t^a\}_{t=1}^{T}$) into visual segment-level logits $\{z_t^v\}_{t=1}^{T} \in \mathbb{R}^{T\times C}$ (*resp.* audio segment-level logits $\{z_t^a\}_{t=1}^{T}$). Segment-level probabilities $\{p_t^v\}_{t=1}^{T}, \{p_t^a\}_{t=1}^{T} \in \mathbb{R}^{T\times C}$ are then obtained by applying the sigmoid function on $\{z_t^v\}_{t=1}^{T}$ and $\{z_t^a\}_{t=1}^{T}$.

Since, only video-level labels y are available during training, Tian et al. [35] introduce an attentive MMIL pooling module to learn to predict video-level probabilities $p \in \mathbb{R}^C$:

$$W_{modal}^{v,a} = \text{Softmax}_{modal}(\text{FC}_{modal}(\tilde{F}^{v,a})), \quad (3)$$

$$W_{time}^{v,a} = \text{Softmax}_{time}(\text{FC}_{time}(\tilde{F}^{v,a})), \quad (4)$$

where FC_{modal} and FC_{time} are two learnable fully-connected layers, $\tilde{F}^{v,a} = \text{Stack}(\tilde{F}^v, \tilde{F}^a) \in \mathbb{R}^{2\times T\times d}$ denotes the stacked visual and audio features along the first dimension, $\text{Softmax}_{modal}(\cdot)$ denotes the softmax operation along the modality dimension (*i.e.*, across v, a), while $\text{Softmax}_{time}(\cdot)$ denotes the softmax operation along the

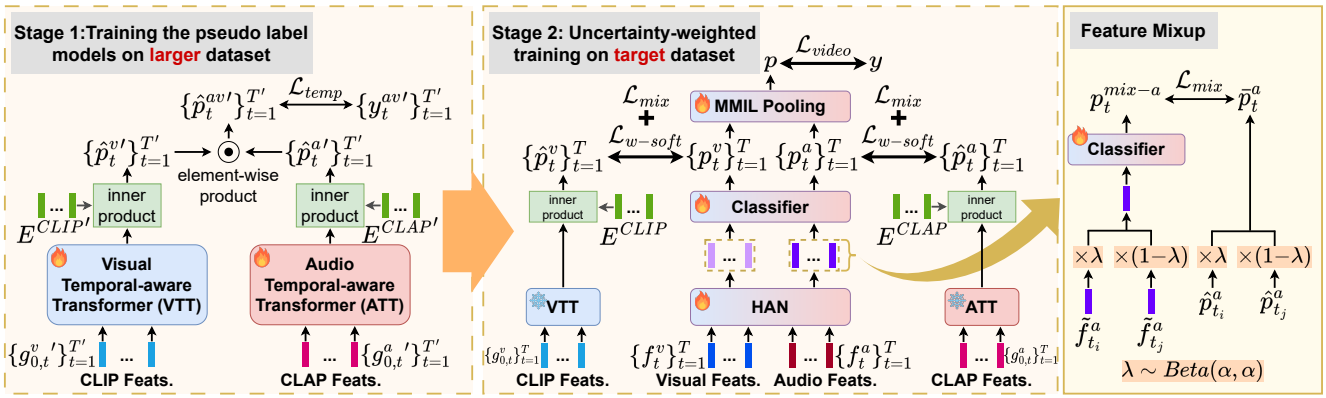

Figure 2. **UWAV framework:** In stage 1, pseudo-label generation modules are equipped with the ability to capture temporal relationships between segments by pre-training on a large-scale, supervised, audio-visual event localization dataset. In stage 2, temporally coherent, uncertainty-weighted pseudo-labels, derived from the pre-trained pseudo-label generation module, are used to guide the learning of the inference model (HAN) aided by a class-balanced loss re-weighting and uncertainty-weighted feature mixup strategy. Note that we use the feature mixup strategy in both modalities while we only show the breakdown of the mixup operation for the audio modality.

temporal dimension (*i.e.*, across $1, \ldots, T$). Video-level probabilities $p \in \mathbb{R}^C$ are then obtained via:

$$p = \sum_{m=\{v,a\}} \sum_{t=1}^{T} W_{modal}^{m,t} \odot W_{time}^{m,t} \odot p_t^m, \tag{5}$$

where $\odot$ denotes the element-wise product. The HAN model is then optimized with the binary cross-entropy (BCE) loss between the estimated video-level probabilities p and video-level labels y: $\mathcal{L}_{video} = \text{BCE}(p, y)$.

4. Proposed Approach

In this section, we detail our proposed approach (UWAV). At a high level, UWAV works by generating better segment-level pseudo-labels to improve the training of a multi-modal transformer-based inference module, *e.g.* HAN. Moreover, UWAV factors in the uncertainty associated with these pseudo-labels, addresses the imbalance in the training data, and introduces self-supervised regularization constraints, which all lead to better performance. Figure 2 shows an overview of our proposed framework.

4.1. Temporally-Coherent Pseudo-Label Synthesis

One major issue that plagues prior works, based on pseudo-label generation [9, 17, 50], is that they do not capture the temporal dependencies between neighboring segments when generating the pseudo-labels. That is, the generated pseudo-labels are not temporally coherent. To plug this void, we propose to incorporate transformer modules [38] into the pseudo-label generation pipeline, which maps CLIP/CLAP encodings of a segment's visual frame/audio to pseudo-labels. Specifically, two separate transformers are introduced, one each for the visual/audio pseudo-label synthesis modules.

Pre-Training: Training transformers often requires sufficiently large training data while datasets commonly used for the weakly-supervised AVVP task are relatively small. To mitigate this challenge, we propose to first pre-train the transformer-equipped pseudo-label generation module on a large-scale, supervised, audio-visual event localization dataset – the UnAV [12] dataset. Specifically, given an audible video of duration T' seconds from the pre-training dataset, we split the video into T' one-second segments $\{V'_t, A'_t\}_{t=1}^{T'}$, with corresponding audio-visual event labels $y_t^{av\prime} \in \{0,1\}^{C'}$, where 1 indicates the presence of an event in both modalities and 0 its absence in at least one modality, while C' denotes the total number of event classes in the pre-training dataset. Next, the video frame at the temporal center of the visual segment is transformed into visual features $G_0^{v\prime} = \{g_{0,t}^{v}{}'\}_{t=1}^{T'} \in \mathbb{R}^{T' \times d_1}$ with CLIP's [28] image encoder. These features are then fed into the corresponding transformer of the visual stream, consisting of L encoder blocks, each block containing a self-attention layer, LayerNorm [1] (LN), and a 2-layer feed-forward network (FFN):

$$\tilde{G}_l^{v\prime} = \text{LN}(G_l^{v\prime} + \text{Attn}(G_l^{v\prime}, G_l^{v\prime}, G_l^{v\prime})), \tag{6}$$

$$G_{l+1}^{v}{}' = \text{LN}(\tilde{G}_l^{v\prime} + \text{FFN}(\tilde{G}_l^{v\prime})). \tag{7}$$

Concurrently, we convert each event category label in the pre-training dataset into a textual event feature $e_c^{CLIP\prime} \in \mathbb{R}^{d_1}$ by filling in the pre-defined caption template: "A photo of <EVENT NAME>" with the corresponding event name and passing it to CLIP's text encoder. Equipped with the visual segment-level features $G_L^{v}{}' = \{g_{L,t}^{v}{}'\}_{t=1}^{T'} \in \mathbb{R}^{T' \times d_1}$ and the textual event features $E^{CLIP\prime} = \{e_c^{CLIP\prime}\}_{c=1}^{C'} \in \mathbb{R}^{C' \times d_1}$, we derive visual segment-level logits $\hat{z}_t^{v\prime} \in \mathbb{R}^{C'}$ and probabilities $\hat{p}_t^{v\prime}$ as follows:

$$\hat{p}_t^{v\prime} = \text{Sigmoid}(\hat{z}_t^{v\prime}), \; \hat{z}_t^{v\prime} = E^{CLIP\prime} \cdot g_{L,t}^{v}{}'^{\top}. \tag{8}$$

Similar operations are performed in the audio pseudo-label generation pipeline. The raw waveforms corresponding to the 1-second audio segments are transformed into audio features $G_0^{a\prime} \in \mathbb{R}^{T' \times d_2}$ with CLAP's [12] audio encoder and fed into the corresponding transformer consisting of L encoder blocks. Correspondingly, the textual event features $E^{CLAP\prime} \in \mathbb{R}^{C' \times d_2}$ are generated with the caption template: "This is the sound of <EVENT NAME>" by passing it through CLAP's text encoder. Audio segment-level logits $\hat{z}_t^{a\prime} \in \mathbb{R}^{C'}$ and probabilities $\hat{p}_t^{a\prime}$ can then be derived in the same manner: $\hat{p}_t^{a\prime} = \text{Sigmoid}(\hat{z}_t^{a\prime})$, $\hat{z}_t^{a\prime} = E^{CLAP\prime} \cdot g_t^{a\prime\top}$.

Since the events occurring in the pre-training dataset (UnAV) are audio-visual, we multiply the segment-level visual and audio event probabilities to enforce the predicted labels to be multi-modal in nature: $\{\hat{p}_t^{av\prime}\}_{t=1}^{T'} \in \mathbb{R}^{T' \times C'}$. This network is then trained with the binary cross-entropy (BCE) loss:

$$\mathcal{L}_{temp} = \text{BCE}(\hat{p}_t^{av\prime}, y_t^{av\prime}),\ \hat{p}_t^{av\prime} = \hat{p}_t^{v\prime} \odot \hat{p}_t^{a\prime}. \quad (9)$$

Pseudo-Label Generation on Target Dataset: With the pre-trained pseudo-label generation modules in place, we proceed to employ them for the pseudo-label generation process in the target dataset for the AVVP task. Specifically, the center frame of each of the visual segments $\{V_t\}_{t=1}^{T}$ of the target dataset are passed into CLIP's image encoder, whose output is passed into the pre-trained visual transformer to generate segment features $G_L^v = \{g_{L,t}^v\}_{t=1}^{T} \in \mathbb{R}^{T \times d_1}$. At the same time, the caption template: "A photo of <EVENT NAME>" is used to obtain textual features corresponding to each of the event classes in the target dataset for the AVVP task: $E^{CLIP} \in \mathbb{R}^{C \times d_1}$. Segment-level visual logits $\hat{z}_t^v \in \mathbb{R}^C$ can be derived by computing their inner product. We also pre-define class-wise visual thresholds $\theta^v \in \mathbb{R}^C$ to transform segment-level visual logits into binary pseudo-labels $\hat{y}_t^v \in \mathbb{R}^C$:

$$\hat{y}_t^v = \mathbb{1}_{\{\hat{z}_t^v > \theta^v\}} \odot y,\ \hat{z}_t^v = E^{CLIP} \cdot g_{L,t}^{v\top}, \quad (10)$$

where y denotes the ground-truth video-level labels, $\mathbb{1}_{\{\cdot\}}$ is the indicator function which returns a value of 1 when the condition is true otherwise 0, and $\odot$ denotes the element-wise product operation. The $\odot$ operation zeroes out the predictions of event classes absent in the video-level label.

A similar pseudo-label generation process is employed on the acoustic side. Raw waveforms of audio segments are first fed into CLAP's audio encoder and then into the pre-trained audio transformer. The event names of the classes in the target dataset for the AVVP task are filled in the caption template: "This is the sound of <EVENT NAME>" to generate textual event features: $E^{CLAP} \in \mathbb{R}^{C \times d_1}$. Segment-level audio logits $\hat{z}_t^a \in \mathbb{R}^C$ and binary pseudo-labels $\hat{y}_t^a \in \mathbb{R}^C$ are then derived using class-wise thresholds $\theta^a \in \mathbb{R}^C$.

With binary segment-level pseudo-labels for both modalities $\hat{y}_t^v, \hat{y}_t^a$ and the predicted probabilities from the inference module (HAN) $\hat{p}_t^v, \hat{p}_t^a$ in place, the inference module can be trained using the binary cross-entropy loss as shown:

$$\mathcal{L}_{hard} = \text{BCE}(p_t^v, \hat{y}_t^v) + \text{BCE}(p_t^a, \hat{y}_t^a). \quad (11)$$

4.2. Training with Pseudo-Label Uncertainty

While pseudo-labels do provide additional supervision for better training of the inference module, they could potentially be noisy, leading to occasionally incorrect training signals. To ameliorate this problem, we propose an uncertainty-weighted pseudo-label training scheme to improve the robustness of the learning process. Instead of simply training with the binary pseudo-labels $\hat{y}_t^v, \hat{y}_t^a$, we leverage the confidence of the pseudo-label estimation module (associated with the predicted pseudo-label) to weigh the training signal for the inference module. This confidence score serves as a measure of the pseudo-label generation module's uncertainty of its prediction. This may be represented as:

$$\hat{p}_t^v = \text{Sigmoid}(\hat{z}_t^v - \theta^v) \odot y;\ \ \hat{p}_t^a = \text{Sigmoid}(\hat{z}_t^a - \theta^a) \odot y. \quad (12)$$

In other words, considering the visual pseudo-label generation pipeline as an example, the farther the logit $\hat{z}_t^v$ is from the threshold θ^v, whether much lower or higher, the more confident the pseudo-label generation module is about the label it predicts (either approaching 0 or 1). Conversely, the closer the logit is to the threshold, the less the certainty about the correctness of the pseudo-labels (probabilities closer to 0.5). An analogous explanation also holds for the audio pseudo-labels. With the uncertainty-weighted pseudo-labels in place, the inference module (HAN) can be trained with the following uncertainty-weighted pseudo-label loss:

$$\mathcal{L}_{soft} = \text{BCE}(p_t^v, \hat{p}_t^v) + \text{BCE}(p_t^a, \hat{p}_t^a). \quad (13)$$

4.3. Uncertainty-weighted Feature Mixup

Due to the lack of full supervision for the weakly-supervised AVVP task, we explore the efficacy of additional regularization via self-supervision to help the models generalize better. Towards this end, prior pseudo-label generation approaches [26, 39] often employ contrastive learning as a tool to better train the inference module. However, due to the inherent noise in the estimated pseudo-labels, positive samples and negative samples may be mislabeled, decreasing the effectiveness of the self-supervisory training. As an alternative, in this work, we explore the effectiveness of feature mixing, as a self-supervisory training signal for additional regularization. In this setting, we mixup the estimated features of any two segments, additively, and train the model to predict the union of the labels of the two segments. However, since the labels in our setting are noisy, the mixed feature is assigned a label derived from a weighted sum of

the uncertainty-weighted pseudo-labels of each of the two segment features. This is illustrated below:

$$\bar{f}^v_{t_i,t_j} = \lambda \tilde{f}^v_{t_i} + (1-\lambda)\tilde{f}^v_{t_j}, \;\; \bar{p}^v_{t_i,t_j} = \lambda \hat{p}^v_{t_i} + (1-\lambda)\hat{p}^v_{t_j} \quad (14)$$

$$\bar{f}^a_{t_i,t_j} = \lambda \tilde{f}^a_{t_i} + (1-\lambda)\tilde{f}^a_{t_j}, \;\; \bar{p}^a_{t_i,t_j} = \lambda \hat{p}^a_{t_i} + (1-\lambda)\hat{p}^a_{t_j}, \quad (15)$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ and α is a hyper-parameter controlling the Beta distribution, and t_i and t_j indicate two segment indices in a batch of video segments.

After mixing the uni-modal segment-level features, we pass them through the classifier of the inference module and apply the sigmoid function to the output, obtaining mixed segment-level event probabilities p_t^{mix-v} and p_t^{mix-a}. These are used to train the inference model with the uncertainty-aware mixup loss, as shown below:

$$\mathcal{L}_{mix} = \text{BCE}(p_t^{mix-v}, \bar{p}_t^v) + \text{BCE}(p_t^{mix-a}, \bar{p}_t^a). \quad (16)$$

4.4. Class-balanced Loss Re-weighting

Besides the aforementioned challenges of the AVVP task, most of the events in the event set are absent in the pseudo-labels of any (segment of a) video (*i.e.*, most event classes are negative events) and only a few events are present (*i.e.*, positive events are much fewer in number). As a result, the model is dominated by the loss from the negative events. When trained without factoring in this bias, the classifier tends to overfit the negative labels and ignore the positive ones. To address this class imbalance issue, we introduce a *class-balanced loss re-weighting* strategy to re-balance the importance of the losses from the negative and positive events for the uncertainty-weighted pseudo-label loss. Specifically, the loss from the positive events is multiplied by a weight proportional to the frequency of the segments with the negative events in the pseudo-labels, while the loss from the negative events is multiplied by a weight proportional to the frequency of the segments with the positive events in the pseudo-labels, as shown below:

$$\mathcal{L}_{w-soft} = \sum_{m\in\{v,a\}} w^m_{pos} \cdot y \cdot \text{BCE}(p^m_t, \hat{p}^m_t) + w^m_{neg} \cdot (1-y) \cdot \text{BCE}(p^m_t, \hat{p}^m_t), \quad (17)$$

$$w^m_{pos} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}(1-\hat{y}^m_{i,t,c})}{NTC} \times W, \quad (18)$$

$$w^m_{neg} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}\hat{y}^m_{i,t,c}}{NTC}, \quad (19)$$

where N denotes the number of videos in the training set, and W is a hyper-parameter.

In summary, the inference module is trained on the AVVP task with the proposed class-balanced re-weighting, applied to the uncertainty-weighted classification loss, and the uncertainty-weighted feature mixup loss, as shown below:

$$\mathcal{L}_{total} = \mathcal{L}_{w-soft} + \mathcal{L}_{mix} + \mathcal{L}_{video}. \quad (20)$$

5. Experiments

We assess the performance of UWAV empirically across two challenging datasets and report its performance, comparing it with existing state-of-the-art approaches both quantitatively and qualitatively. Additionally, through multiple ablation studies, we bring out the effectiveness of the different elements of our proposed approach and the choices of different hyper-parameters. For additional details, ablation studies, and more qualitative results, we refer the reader to our supplementary material.

5.1. Experimental Setup

Datasets: We evaluate all competing methods on the *Look, Listen, and Parse* (LLP) dataset [35], which is the principal benchmark dataset for the AVVP task. The dataset consists of $11,849$ video clips sourced from YouTube. Each clip is 10 seconds long and represents one or more of 25 diverse event categories, such as human activities, animals, musical instruments, and vehicles. The dataset is split into training, validation, and testing sets, following the official split [35]: $10,000$ videos for training, 649 videos for validation, and $1,200$ videos for testing. While the training set of this dataset is only associated with video-level labels of the events, the validation and testing split is labeled with segment-level event labels for evaluation purposes. Additionally, to demonstrate the generalizability of our method, we conduct a similar study on the *Audio Visual Event* (AVE) recognition dataset [34]. The AVE dataset consists of $4,143$ video clips crawled from YouTube, each 10 seconds long. It is split into $3,339$ videos for training, 402 for validation, and 402 for testing. It includes 29 event categories encompassing human activities, animals, musical instruments, vehicles, and a "background" class (*i.e.*, no event occurs). Unlike the LLP dataset, each video in the AVE dataset contains only one audio-visual event. Here too, the training data is only provided with video-level labels while the validation and test splits are annotated with ground-truth event labels for each one-second segment, which either is a specific audio-visual event or "background".

Metrics: For the LLP dataset, following the standard evaluation protocol [35], all models are evaluated using macro F1-scores calculated for the following event types: (i) audio-only (A), (ii) visual-only (V), and (iii) audio-visual (AV). Type@AV (Type) and Event@AV (Event) are two additional metrics that evaluate the overall performance of the model, where Type@AV is the mean of the F1-scores for the A, V, and AV events, while Event@AV is the F1-score of all events regardless of the modality in which they occur. Evaluations are conducted at both the segment-level and the event-level. At the segment-level, the model's predictions are compared with the ground truth on a per-segment basis. At the event-level, consecutive positive segments for

Table 1. **Comparison with state-of-the-arts methods on the LLP dataset.** Best performances are in **bold**, second-best in underlined.

Method	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
HAN [35]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MA [39]	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6
JoMoLD [8]	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5
CMPAE [10]	64.2	66.4	59.2	63.3	62.8	56.6	63.7	51.8	57.4	55.7
PoiBin [27]	63.1	63.5	57.7	61.4	60.6	54.1	60.3	51.5	55.2	52.3
VPLAN [50]	60.5	64.8	58.3	61.2	59.4	51.4	61.5	51.2	54.7	50.8
VALOR [17]	61.8	65.9	58.4	62.0	61.5	55.4	62.6	52.2	56.7	54.2
LSLD [9]	62.7	67.1	59.4	63.1	62.2	55.7	64.3	52.6	57.6	55.2
PPL [26]	**65.9**	66.7	61.9	64.8	63.7	57.3	64.3	54.3	59.9	**57.9**
CoLeaf [31]	64.2	64.4	59.3	62.6	62.5	57.6	63.2	54.2	57.9	55.6
LEAP [51]	62.7	65.6	59.3	62.5	61.8	56.4	63.1	54.1	57.8	55.0
UWAV (Ours)	64.2	**70.0**	**63.4**	**65.9**	**63.9**	**58.6**	**66.7**	**57.5**	**60.9**	57.4

the same event are grouped together as a single event. The F1-score is then computed using a mIoU threshold of 0.5. For the AVE dataset, we follow Tian et al. [34] and use accuracy as the evaluation metric. An event prediction of a segment is considered correct if it matches the ground-truth label for that segment.

Implementation Details: In line with prior work [35], each 10-second video in both datasets is split into 10 segments of one second each, where each segment contains 8 frames. For the LLP dataset, pre-trained ResNet-152 [13] and R(2+1)D-18 [37] are used to extract 2D and 3D visual features, respectively. The pre-trained VGGish [14] network is used to extract features from the audio, sampled at 16KHz. For the AVE dataset however, akin to prior work [17], we extract visual features from pre-trained CLIP and R(2+1)D while CLAP is used to embed the audio stream. For both datasets, we set the number of encoder blocks L in the temporal-aware model to 5, α for the Beta distribution in the feature mixup strategy to 1.7, and W in the class-balanced loss re-weighting step to 0.5. The pseudo-label generation modules and the inference model (HAN) are trained with the AdamW optimizer [20]. For improved performance on the AVE dataset, we replace $\hat{p}^a_{t_i}, \hat{p}^a_{t_j}$ in Eq. 15 with $\lceil \hat{p}^a_{t_i} \rceil, \lceil \hat{y}^a_{t_j} \rceil$ and make corresponding modifications in the visual counterparts as well.

Baselines: We demonstrate the effectiveness of UWAV by comparing against an extensive set of baselines. For the LLP dataset, this includes video-level pseudo-label generation methods (MA [39], JoMoLD [8]), segment-level pseudo-label generation methods (VALOR [17], LSLD [9], PPL [26]), and the recently released works (CoLeaf [31], LEAP [51]). On the other hand, for the AVE dataset, baseline approaches with publicly available implementation, which use the state-of-the-art feature backbones (akin to ours), such as HAN [35] and VALOR [17] are used.

5.2. Results

Comparison with Previous Methods on LLP: As shown in Table 1, UWAV surpasses previous methods, across almost all metrics. Notably, we achieve an F-score of **70.0** on the segment-level visual event, **65.9** on the segment-level Type@AV, and **66.7** on the event-level visual event metric. This corresponds to a gain of 1.1% on segment-level Type@AV F-score and a 1% improvement on event-level Type@AV F-score, over our closest competitor PPL [26]. Of particular note, is the fact that our segment-level and event-level F-scores improve by more than 3%, over PPL, for visual events. When compared to other recently published works, such as VALOR [17], CoLeaf [31] and LEAP [51], UWAV outperforms them by up to 3% on both segment and event-level Type@AV F-score while gains on visual events are up to 5% on segment-level F-score.

These observations are consistent with our qualitative comparisons, as well. In the example on the left of Figure 3, our model successfully recognizes and temporally localizes the lawn mower event visually, whereas VALOR [17] (a recent state-of-the-art approach with publicly available implementation) misclassifies it as a chainsaw. Additionally, our model also accurately localizes the intermittent sound of the lawn mower. In contrast, VALOR not only misclassifies the sound of the lawn mower as that of a chainsaw but also incorrectly predicts that someone is talking in the video. In the example on the right of Figure 3, our model does not err in recognizing either the visual presence or the audio presence of the telephone, while VALOR fails to accurately predict the events in either modality.

Comparison with Previous Methods on AVE: To demonstrate the generalizability of our method, we evaluate UWAV on the AVE [34] dataset and compare its per-

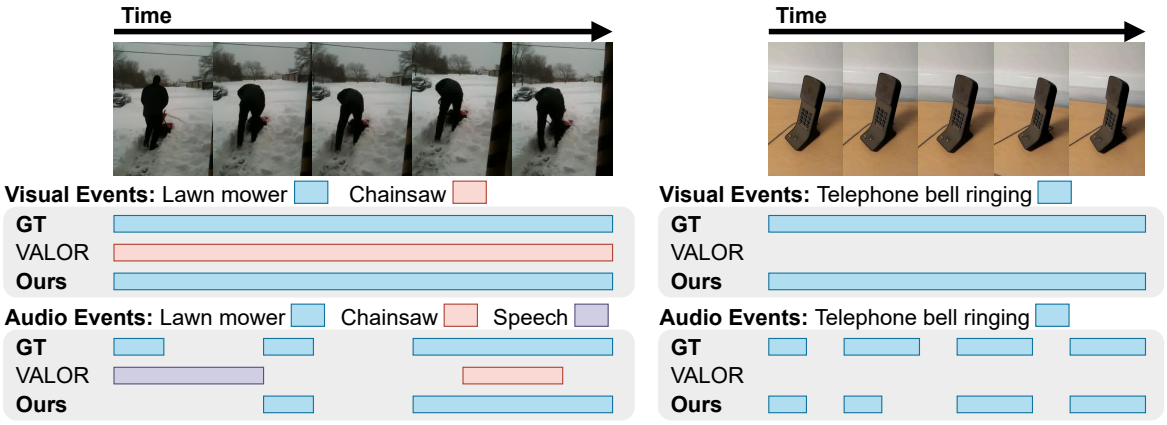

Figure 3. **Comparison between predictions by UWAV and competing AVVP methods on the LLP dataset.** "GT": ground truth.

Table 2. **Model performances on the AVE dataset.** CLIP, R(2+1)D-18, and CLAP are used as feature backbones.

Method	HAN [35]	VALOR [17]	UWAV (Ours)
Acc.(%)	75.3	80.4	**80.6**

Table 3. **Accuracy of the generated pseudo-labels on LLP.**

Method	Segment-level				
	A	**V**	**AV**	**Type**	**Event**
VALOR [17]	**80.5**	61.7	55.7	66.0	74.6
PPL [26]	61.7	61.8	57.5	60.6	59.4
UWAV (Ours)	78.4	**74.5**	**65.5**	**72.8**	**78.4**

formance with that of previous works. From Table 2, we observe that with the same backbone features, UWAV surpasses VALOR, our closest competitor, even on this small-scale dataset.

Accuracy of the Generated Pseudo-Labels: To evaluate the efficacy of our pseudo-label generation pipeline, we compare the accuracy of our generated pseudo-labels against those obtained from competing methods (with publicly available implementation) [17, 26] on the test set of the LLP dataset. As shown in Table 3, our pre-trained temporally-dependent pseudo-label generation scheme generates more accurate segment-level pseudo-labels than previous methods, by up to 6% on the segment-level Type@AV F-score, attesting to the advantages of factoring in inter-segment temporal dependencies.

5.3. Ablation Study

To demonstrate the potency of the different elements of our proposed method, UWAV, we conduct ablation studies. In particular, the proposed uncertainty-weighted pseudo-label based training, the uncertainty-weighted feature mixup scheme, and the class-balanced loss re-weighting schemes are ablated. As shown in Table 4, incorporating the uncertainty-weighted pseudo-label training step improves

Table 4. **Ablation study of the proposed components in UWAV.** "Binary" denotes training with binary pseudo-labels. "Soft" denotes training with uncertainty-weighted pseudo-labels.

Binary	Soft	Re-weight	Mixup	Segment-level				
				A	**V**	**AV**	**Type**	**Event**
✓				62.7	67.7	61.9	64.2	62.2
	✓			63.0	68.3	61.8	64.4	62.8
	✓	✓		63.6	69.5	63.0	65.4	63.1
	✓		✓	63.9	69.0	62.8	65.2	63.4
	✓	✓	✓	**64.2**	**70.0**	**63.4**	**65.9**	**63.9**

the segment-level Type@AV F-score by 2%, compared to using binary pseudo-labels. This demonstrates the benefit of accounting for the uncertainty in the pseudo-label estimation module. Moreover, sans the class-balanced loss re-weighting strategy, the model's performance is worse off by 1% on the Type@AV F-score, revealing the erroneous bias in the model's prediction arising from a skew of the class distribution. On the other hand, introducing the uncertainty-weighted feature mixup results in a gain of 0.8% on the Type@AV F-score, which underscores the importance of this self-supervised regularization.

6. Conclusions

In this work, we address the challenging task of weakly-supervised audio-visual video parsing (AVVP), which presents a two-fold challenge: (i) potential misalignment between the events of the audio and visual streams, and (ii) the lack of fine-grained labels for each modality. We observe that by considering the temporal relationship between segments, our proposed method (UWAV) is able to provide more reliable pseudo-labels for better training of the inference module. In addition, by factoring in the uncertainty associated with these estimated pseudo-labels, regularizing the training process with a feature mixup strategy, and correcting for class imbalance, UWAV achieves state-of-the-art results on the LLP and AVE datasets.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *ICCV*, 2021. 2

[3] Moitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. Learning audio-visual dynamics using scene graphs for audio source separation. In *NeurIPS*, 2022. 1

[4] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 2

[5] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, 2021.

[6] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *ICLR*, 2021. 2

[7] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *CVPR*, 2023. 1, 2

[8] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *ECCV*, 2022. 2, 3, 7

[9] Yingying Fan, Yu Wu, Yutian Lin, and Bo Du. Revisit weakly-supervised audio-visual video parsing from the language perspective. In *NeurIPS*, 2023. 2, 3, 4, 7

[10] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *CVPR*, 2023. 7

[11] Shijie Geng, Peng Gao, Moitreya Chatterjee, Chiori Hori, Jonathan Le Roux, Yongfeng Zhang, Hongsheng Li, and Anoop Cherian. Dynamic graph representation learning for video dialog via multi-modal shuffled transformers. In *AAAI*, 2021. 2

[12] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *CVPR*, 2023. 2, 4, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[14] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017. 7

[15] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *CVPR*, 2022. 2

[16] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *CVPR*, 2023. 2

[17] Yung-Hsuan Lai, Yen-Chun Chen, and Frank Wang. Modality-independent teachers meet weakly-supervised audio-visual event parser. In *NeurIPS*, 2023. 2, 3, 4, 7, 8

[18] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 2022. 2

[19] Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, and Yanfeng Wang. Audio-visual segmentation via unlabeled frame exploitation. In *CVPR*, 2024. 1, 2

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7

[21] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *ICCV*, 2021. 2

[22] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *ICCV*, 2023. 2

[23] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *NeurIPS*, 2022. 2, 3

[24] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *CVPR*, 2023. 1, 2

[25] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In *WACV*, 2022. 2

[26] Kranthi Kumar Rachavarapu, Kalyan Ramakrishnan, et al. Weakly-supervised audio-visual video parsing with prototype-based pseudo-labeling. In *CVPR*, 2024. 2, 3, 5, 7, 8

[27] Kranthi Kumar Rachavarapu et al. Boosting positive segments for weakly-supervised audio-visual video parsing. In *ICCV*, 2023. 7

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 4

[29] Varshanth Rao, Md Ibrahim Khalil, Haoda Li, Peng Dai, and Juwei Lu. Dual perspective network for audio-visual event localization. In *ECCV*, 2022. 1, 2

[30] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023. 2

[31] Faegheh Sardari, Armin Mustafa, Philip JB Jackson, and Adrian Hilton. Coleaf: A contrastive-collaborative learning framework for weakly supervised audio-visual video parsing. In *ECCV*, 2024. 7

[32] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *ICCV*, 2023. 1, 2

[33] Ankit Shah, Shijie Geng, Peng Gao, Anoop Cherian, Takaaki Hori, Tim K Marks, Jonathan Le Roux, and Chiori Hori. Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning. In *ICASSP*, 2022. 2

[34] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1, 2, 6, 7

[35] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020. 1, 2, 3, 6, 7, 8

[36] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *CVPR*, 2021. 2

[37] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 7

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3, 4

[39] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, 2021. 2, 3, 5, 7

[40] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023. 2, 3

[41] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024. 2

[42] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *CVPR*, 2024. 2

[43] Yuxin Ye, Wenming Yang, and Yapeng Tian. Lavss: Location-guided audio-visual spatial audio separation. In *WACV*, 2024. 1, 2

[44] Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *IEEE Robotics and Automation Letters*, 2023. 2

[45] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *ACM MM*, 2022. 2, 3

[46] Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. Sound adversarial audio-visual navigation. In *ICLR*, 2022. 2

[47] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *ICCV*, 2021. 2

[48] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *CVPR*, 2021. 2

[49] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *ECCV*, 2022. 1, 2

[50] Jinxing Zhou, Dan Guo, Yiran Zhong, and Meng Wang. Improving audio-visual video parsing with pseudo visual labels. *arXiv preprint arXiv:2303.02344*, 2023. 2, 3, 4, 7

[51] Jinxing Zhou, Dan Guo, Yuxin Mao, Yiran Zhong, Xiaojun Chang, and Meng Wang. Label-anticipated event disentanglement for audio-visual video parsing. *arXiv preprint arXiv:2407.08126*, 2024. 2, 3, 7