Reinforced Label Denoising for Weakly-Supervised Audio-Visual Video Parsing

Yongbiao Gao[1,2,3,5], Xiangcheng Sun[1,2], Guohua Lv[1,2], Deng Yu[4], Sijiu Niu[5]

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education,

Shandong Computer Science Center (National Supercomputer Center in Jinan),

Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[2]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing,

Shandong Fundamental Research Center for Computer Science, Jinan, China

[3]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Application (Southeast University),

Ministry of Education, China

[4]School of Artificial Intelligence, Shandong University

[5] Shandong Key Laboratory of Ubiquitous Intelligent Computing, Jinan, China

Abstract

Audio-visual video parsing (AVVP) aims to recognize audio and visual event labels with precise temporal boundaries, which is quite challenging since audio or visual modality might include only one event label with only the overall video labels available. Existing label denoising models often treat the denoising process as a separate preprocessing step, leading to a disconnect between label denoising and AVVP tasks. To bridge this gap, we present a novel joint reinforcement learning-based label denoising approach (RLLD). This approach enables simultaneous training of both label denoising and video parsing models through a joint optimization strategy. We introduce a novel AVVP-validation and soft inter-reward feedback mechanism that directly guides the learning of label denoising policy. Extensive experiments on AVVP tasks demonstrate the superior performance of our proposed method compared to label denoising techniques. Furthermore, by incorporating our label denoising method into other AVVP models, we find that it can further enhance parsing results.

Keywords: Video Parsing, Reinforcement Learning, Label Denoising, Weakly-Supervised.

1. Introduction

With the rapid progress of Internet technology, large-scale visual data can be found on the Internet, bringing significant opportunities for novel processing of visual information, as well as commercial applications. The Computational Visual Media Conference series, of which this is to be the first conference, is intended to provide a major new international forum for exchanging novel research ideas and significant practical results both underpinning and applying Visual Media. The primary rationale for this new conference series is to target cross disciplinary research which amalgamates aspects of computer graphics, computer vision, machine learning, image processing, video processing, visualization and geometric computing. Original research is sought in areas concerning the classification, composition, retrieval, synthesis, and understanding of visual media.

The way humans perceive knowledge is through multiple modalities, including auditory, visual, tactile, and taste. Numerous studies have demonstrated that integrating these various senses can enhance natural knowledge perception [17, 29, 26, 4]. Specifically, visual and auditory are particularly prevalent in conveying vast amounts of information during interactions with the environment.

In recent years, the majority of research focus on visual scene understanding [15, 32] or video parsing [24, 23], often overlooking other modalities. Prior works in multimodal learning [3, 38] primarily examine the synergy between text and visual modalities to acquire joint representations. However, auditory and visual data often coexist and provide complementary cues in human perception. Therefore, more recently, some works [31, 40, 5, 44, 25, 27, 18] investigate on localizing audio and visual event tasks, called Audio-Visual Video Parsing (AVVP), which aims to generate temporal audio, video, and audio-visual event labels. Most of existing works explore weakly-supervised learning for AVVP task since densely annotated event modality and category label with temporal boundaries is extremely expensive and time-consuming. For instance, [31] formulates the AVVP task as a Multimodal Multiple Instance Learning (MMIL) problem and introduces an attentive MMIL pooling method to adaptively explore useful audio and visual content from different temporal extents and modali-
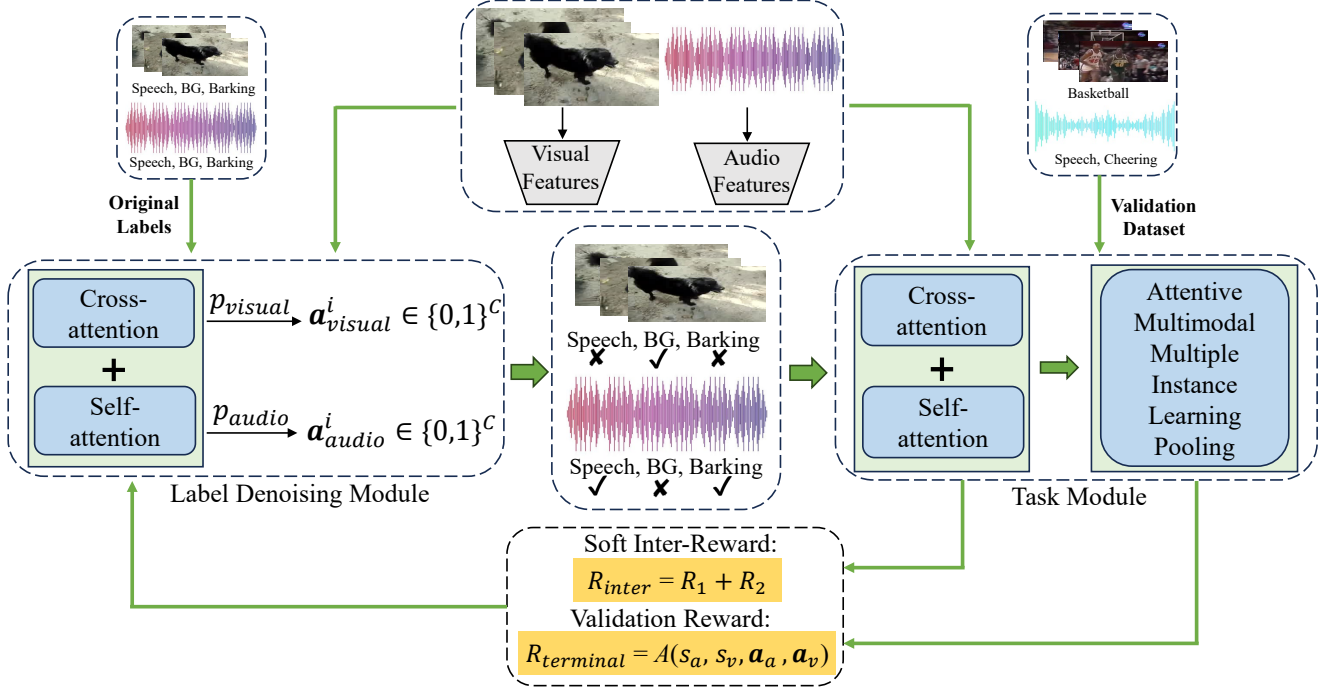
Figure 1. 1) Feature extraction is performed on both the image and speech data. 2) The noisy labels, along with the extracted features from the two modal models, are input into a label denoising model to make decisions about label noise. 3) The denoised labels are then used to train a video parsing model, which is validated on a validation set to obtain quantitative analysis results of the parsing. 4) Based on the obtained quantitative analysis results, the reward is calculated. The obtained reward is fed back to the label denoising model to guide its training process. $\mathbf{a}^i_{visual}$ and $\mathbf{a}^i_{audio}$ are actions of the denoising model.

ties. While existing audio-visual parsing works often assume temporal alignment of audio and visual events, this alignment hypothesis may not hold in practical scenarios. As pointed out by [37, 6], "some event clues that do not appear in one of the two modalities." This is exemplified by the saying "I have not seen the person, but I have heard his voice." For example, in a basketball game video, even if the commentator cannot be seen, their voice and cheering can be heard, making sound signals crucial for comprehensive video understanding. Thus, the commentator's speech serves as a noisy label for the visual modality, and the basketball is considered audio-specific noise in some video clips. This issue is defined as **modality-specific noisy**.

Pioneering researchers in the field of audio-visual video parsing have made concerted efforts to address the modality-specific noisy problem. For example, an audio-visual modality exchange method [37] aimed at deriving precise modality-aware event labels is proposed to, thus safeguard models from being misled by ambiguous overall labels. In a similar vein, [6] tackle the modality-specific noisy issue by collaboratively computing a nosie estimator and forward modality losses. Leveraging the estimated noise ratios, they prioritize noisy samples based on forward losses. Notably, both the cross-modal and joint modal exchange mechanisms treat label denoising as a preliminary

stage, disconnected from the subsequent AVVP process. This limitation precludes direct guidance of the denoising process by AVVP outcomes. Recently, VALOR [42] incorporates large-scale contrastively pre-trained models as the modality teachers to enhance learning with weak labels. An iterative optimization approach [27] using the EM algorithm is proposed to overcome the challenge of weak labels. However, the two most recent works ignore the modality-specific noisy problem. Since it is unknown which labels are noise in video parsing, the exploration strategy based on reinforcement learning can be utilized to tackle the problem of label denoising. Therefore, we introduce an innovative training strategy that integrates label denoising and audio-visual video parsing into a unified learning framework. This approach enables seamless interaction between denoising and parsing components, leading to more robust and accurate audio-visual video analysis.

In this paper, we introduce a novel Reinforcement Learning-based Label Denoising (RLLD) framework for audio-visual video parsing. The objective of RLLD is to enhance the identification of noisy labels and facilitate label denoising, thereby improving the overall performance of AVVP. Figure 1 provides an overview of the RLLD framework. The core component of RLLD is a reinforcement learning agent that is trained to identify noisy labels. This

agent is guided by a joint reward signal derived from both the validation results and the soft inter-reward. This approach ensures that the denoising agent can maintain prediction accuracy while reducing convergence time. In addition, RLLD integrates label denoising and AVVP into a unified training framework. This integration enables the direct mitigation of label denoising and AVVP, facilitating more effective learning and optimization.

RLLD not only directly identifies noisy labels from a single modality, eliminating the need for computationally expensive cross-modal comparison, but also trains the denoising agent using validation results from the AVVP model. This approach ensures a more direct impact on enhancing AVVP performance, leading to superior temporal localization outcomes. We validate the effectiveness of RLLD on the standard AVVP dataset under several settings. The results demonstrate that RLLD achieves better performance than existing state-of-the-art label denoising models and further enhances the performance of state-of-the-art AVVP models. To summarize, our key contributions are as follows:

- We propose to address the label denoising issue in AVVP by reinforcement learning to learn the noisy label discovery strategy. To the best of our knowledge, we are the first to introduce reinforcement learning into AVVP to solve the label denoising problem.

- We carefully design the validation-based reward and the soft inter-reward functions to guide the label denoising strategy learning, which directly drives the label denoising for improving the performance of AVVP.

- The experiments on the AVVP dataset validate the effectiveness of our method, which can alleviate modality-specific noise during the jointly training framework.

2. Related Work

The traditional audio-visual learning primarily emphasize the representation learning [1, 19, 2, 7, 16], video captioning [28, 20, 39], audio-visual sound separation [34, 8, 11], *etc*. For example, [2] propose a knowledge distillation-based procedure to transfer discriminative visual-knowledge from well established visual recognition models into the sound modality to get the better sound representation. Vid2Seq [39] leverages unlabeled narrated videos for dense video captioning by reformulating sentence boundaries of transcribed speech as pseudo event boundaries, and using the transcribed speech sentences as pseudo event captions. All the past audio-visual related works assume audio and visual data are always correlated. In practice, the event label is contained in different modalities with different boundaries. Therefore, [31] introduce

a new fundamental problem: *audio-visual video parsing (AVVP)* that recognizes every event label containing in audio and visual modalities with temporal boundaries.

HAN [31] is the pioneering weakly supervised framework to solve the audio-visual video parsing problem. HAN formulates the AVVP as a multi-instance learning problem and use a hybrid attention network to learn the unimodal and the cross-modal temporal contexts. The label smoothing technique is adopted to suppress the impact of noisy label from individual modality. Furthermore, MM-Pyramid [40] attempt to capture temporal pyramid features via several stacking pyramid units. The previous proposed AVVP models take the overall event labels to supervise both visual and audio modalities learning. Recently, several latest models are proposed to solve the AVVP task like MGN [25], DHHN [18], CVCMS [22], CPSP [43], CMPAE [10], and poiBin [27] etc. For example, audio, visual, and audio-visual streams are treated separately and an audio-visual class co-occurrence module is adopted to jointly explores the relationship of different categories among all streams [22]. CPSP [43] introduces the available full or weak label as a prior that constructs the exact positive-negative samples for contrastive learning. In practice, visual modality may not include the audio event, and vice versa. Therefore, the denoising methods [37, 6] aim to remove the noisy label for each modality by considering cross-modal connections and differences. LSLD [9] designs language prompts to describe all cases of event appearance for each video since the language could freely describe how various events appear in each segment beyond fixed labels. Zhou *et al.*[45] use CLIP and CLAP to estimate the events in each video segment and generate segment-level visual and audio pseudo labels, respectively. However, all the label denoising approaches in AVVP formulate the modality-specific denoising as a preprocessing stage, leading to a disconnect between label denoising and AVVP tasks.

In contrast to these methods, we generate reliable event labels independently for each modality via a joint training framework. We utilize validation results from the AVVP model to guide the learning of the label denoising strategy. This approach directly addresses the label denoising problem within the AVVP learning process, enabling more effective modality-specific denoising and improved overall performance.

3. Method

In this section, we elaborate on the new framework, Reinforcement Learning-based Label Denoising (RLLD) for AVVP. As depicted in Figure 1, RLLD comprises two distinct networks: label denoising network and task network. The label denoising network is implemented as a hybrid attention network for generating policy to determine which modality-specific label is the noisy label. We use the HAN

[31] as the task network to guide the denoising policy learning, though our method can not only adapt to HAN but also to any task networks for AVVP, such as [42, 27, 25, 18], among others.

3.1. Problem Statement

The input video sequence is denoted as $\{A_t, V_t\}_{t=1}^T$, where T represents the overall duration of the video. Each A_t and V_t corresponds to the audio and visual modalities, respectively. The objective of AVVP is to segment the boundaries of audio and visual events and assign event labels that are present in each modality. During the training stage, only the overall video-label $\boldsymbol{y} \in \{0,1\}^C$ can be used to train the parsing network. Therefore, AVVP is a classic multiple instance learning problem. During validation stage, the audio, visual and audio-video event labels $\boldsymbol{y}_a^t \in \{0,1\}^C$, $\boldsymbol{y}_v^t \in \{0,1\}^C$, $\boldsymbol{y}_{av}^t \in \{0,1\}^C$ are available to assess whether the event occurs in both the visual and audio tracks at time t. In our formulation, following MA [37] and JoMoLD [6], we aim to remove the modality-specific noisy label to improve the parsing performance. As a novel framework for label denoising in AVVP, RLLD differs from MA and JoMoLD in that we employ reinforcement learning to integrate label denoising with AVVP into a unified learning framework. This approach allows for joint optimization of denoising policy and AVVP task-specific objective.

3.2. The Label Denoising Network

We formulate the label denoising as a sequential decision-making process. Specifically, we develop a deep label denoising network that predicts probabilities for audio and visual tracks and determines which event label to remove from each modality based on the predicted probability distributions. We present a joint framework for training the label denoising network by using reinforcement learning.

Figure 1 visually illustrates how RLLD identifies noisy labels. A reinforcement learning agent is employed to map actions a_t from the state s_t to remove the event labels for each modality. We argue that the original event labels play a crucial role in identifying noisy labels. Therefore, the reinforcement learning state s_t is defined as a combination of features and original noisy label representation, which is denoted as follows:

$$
\begin{aligned}
s_{audio}^i &= concat(f_{audio}^t, y_a^i), \\
s_{visual}^i &= concat(f_{visual}^t, y_v^i),
\end{aligned}
\tag{1}
$$

where f_a^i and f_v^i are audio and visual features extracted from pre-trained models. i represents the index of the video. After providing the audio and visual track states for a given video, following the approach of HAN [31], at each time step t, self-attention and cross-attention mechanisms are utilized to learn the audio and visual hidden state representations, respectively. This process is formalized as follows:

$$
\hat{h}_{audio}^i = s_{audio}^i + g_{sa}(s_{audio}^i, \boldsymbol{s}_{audio}) + g_{ca}(s_{audio}^i, \boldsymbol{s}_{visual}),
$$
$$
\hat{h}_{visual}^i = s_{visual}^i + g_{sa}(s_{visual}^i, \boldsymbol{s}_{visual}) + g_{ca}(s_{visual}^i, \boldsymbol{s}_{audio}),
\tag{2}
$$

where $\boldsymbol{s}_{audio} = [s_{audio}^1; ...; s_{audio}^T]$ and $\boldsymbol{s}_{visual} = [s_{visual}^1; ...; s_{visual}^T]$. $g_{sa}(\cdot)$ and $g_{ca}(\cdot)$ are defined as follows:

$$
g_{sa}(s_{audio}^t, \boldsymbol{s}_{audio}) = softmax(\frac{s_{audio}^t \boldsymbol{s}_{audio}^{'}}{\sqrt{d}})\boldsymbol{s}_{audio},
$$
$$
g_{ca}(s_{audio}^t, \boldsymbol{s}_{visual}) = softmax(\frac{s_{audio}^t \boldsymbol{s}_{visual}^{'}}{\sqrt{d}})\boldsymbol{s}_{visual},
\tag{3}
$$

where d is the dimension of the audio and visual state. The fully connected (FC) layer, followed by the sigmoid function, is responsible for predicting the probability p^t for each modality. Based on this probability, a modality-removal action $\boldsymbol{a}^t$ is sampled:

$$
\begin{aligned}
p_{audio}^i &= \sigma(W\hat{h}_{audio}^i), \\
p_{visual}^i &= \sigma(W\hat{h}_{visual}^i), \\
\boldsymbol{a}_{audio}^i &\sim Bernoulli(p_{audio}^i), \\
\boldsymbol{a}_{visual}^i &\sim Bernoulli(p_{visual}^i),
\end{aligned}
\tag{4}
$$

where σ represents the sigmoid activate function, $\boldsymbol{a}_{audio}^i \in \{0,1\}^C$ and $\boldsymbol{a}_{visual}^i \in \{0,1\}^C$ sampled from Bernoulli distribution indicate whether the corresponding index event label is removing or not. Therefore, the revised label for each modality is obtained by:

$$
\begin{aligned}
\hat{y}_a^i &= \boldsymbol{a}_{audio}^i \odot y_a^i, \\
\hat{y}_v^i &= \boldsymbol{a}_{visual}^i \odot y_v^i,
\end{aligned}
\tag{5}
$$

where $\odot$ indicates element-wise multiplication. $\hat{y}_a^i$ and $\hat{y}_v^i$ are the event labels for audio and visual tracks after removing noisy labels.

3.3. The Task Network

During training, RLLD will receive a reward $R(S)$ that evaluates the quality of the denoising policy, and the objective of RLLD is to maximize the expected rewards over time by removing the noisy labels for each modality. In general, a high-quality label space is expected to obtain better performance in AVVP task. To this end, we propose a novel reward mechanism that directly feedback from the validation of the AVVP task.

We use the HAN as the basic task network with parameters W. We assume that the audio and visual snippets from temporal aggregated features are denoted as $\{\hat{f}_{audio}^t, \hat{f}_{visual}^t\}$. The aggregated features are predicted from self-attention and cross-attention networks similar to

Eq.(2). Different from the label denoising network, the input of the task network only includes audio and visual features extracted from pre-trained models. The label representation is omitted in the task network. Subsequently, the attentive MMIL pooling is employed to predict video-level event probabilities. We denote the union of $\hat{y}_a^i$ and $\hat{y}_v^i$ as the ground-truth of the video-level label.

$$\hat{y} = \hat{y}_a^i \cup \hat{y}_v^i, \qquad (6)$$

Therefore, the binary cross-entropy loss $CE(\cdot)$ is optimized to train the video-level prediction model.

$$\mathcal{L}_{video} = -\sum_{c=1}^{c} \hat{y}[c] log(\bar{p}[c]), \qquad (7)$$

where $\bar{p}$ is the predicted video-level event labels. Unlike the HAN approach, which employs label smoothing to mitigate modality bias, we utilize the outcomes from the label denoising network as the ground truth for both audio and visual modalities. The learning of audio and visual modalities is guided by the following loss function:

$$\mathcal{L}_{a,v} = -(\sum_{c=1}^{c} \hat{y}_a[c] log(\bar{p}_a[c]) + \sum_{c=1}^{c} \hat{y}_v[c] log(\bar{p}_v[c])), \quad (8)$$

where the $\hat{y}_a$ and $\hat{y}_a$ are predicted audio and visual event labels from attentive MMIL pooling network.

Recognizing that validation accuracy serves as a direct indicator of the quality of the label denoising, we design a soft inter-reward to guide the learning of the label denoising policy at each step. The soft inter-reward is computed using two components:

Kullback-Leibler Divergence-based Reward (R_1): This component measures the divergence between the soft labels and the revised labels, encouraging the denoising policy to produce labels that are closer to the soft labels.

$$R_1 = \exp\left(-\left(\sum_{j=1}^{m} l_{soft}^{(j)} \ln \frac{l_{soft}^{(j)}}{l_t^{in_{(j)}}} + \sum_{j=1}^{m} l_t^{in_{(j)}} \ln \frac{l_t^{in_{(j)}}}{l_{soft}^{(j)}}\right)/2\right) \quad (9)$$

Cosine Similarity-based Reward (R_1): This component calculates the cosine similarity between the soft labels and the revised labels, providing a measure of how well the revised labels align with the soft labels in direction.

$$R_2 = \frac{\sum_{j=1}^{m} l_{soft}^{(j)} l_t^{in_{(j)}}}{\sqrt{\sum_{j=1}^{m} \left(l_{soft}^{(j)}\right)^2} \sqrt{\sum_{j=1}^{m} \left(l_t^{in_{(j)}}\right)^2}} \qquad (10)$$

The soft inter-reward R_{inter} is then a weighted sum of R_1 and R_2

$$R_{inter} = \alpha_1 R_1 + \alpha_2 R_2, \qquad (11)$$

where α_1 and α_2 are hyperparameters that control the relative importance of the two components. And l_{soft} is the soft label from the original labels by a smoothing operation. l_t indicates the revised labels at every time-step. The R_{inter} gives an immediate reward after the each denoising action and provides a direction for the denoising policy learning. In addition, a validation-based reward is carefully designed to directly guides the label denoising for the improvement of AVVP.

$$R_{terminal} = A(s_a, s_v, \boldsymbol{a}_a, \boldsymbol{a}_v) \qquad (12)$$

where $A(\cdot)$ represents the validation result on validation dataset $\mathcal{V}$. It is obtained from the parsing result. In our setting, the $R_{terminal}$ equals the Segment-level@Type. The final reward is combined by the soft inter-reward R_{inter} and the terminal reward $R_{terminal}$.

$$R(S) = R_{terminal} + R_{inter}, \qquad (13)$$

Since R is non-differentiable, we use policy gradient, described in the following section, to update the label denoising network. By utilizing the proposed framework, we successfully integrate label denoising and AVVP into a unified training paradigm. The outcomes of label denoising can be directly utilized to train the AVVP model, while the results of AVVP serve as direct guidance for learning label denoising strategies. This joint training approach enables complementary learning between the two components.

3.4. Training with Policy Gradient

The objective of label denoising for AVVP is to learn a policy function π_θ with parameters θ by maximizing the expected rewards:

$$J(\theta) = \mathbb{E}_{p_\theta}(a_{1:N})[R(S)], \qquad (14)$$

where $a_{1:N}$ denotes the probability distributions over possible actions for each modality, and $R(S)$ is computed by Eq.(9). π_θ is defined by our proposed RLLD.

Due to the fact that rewards can only be obtained after training video parsing verification based on denoising model results, there is a problem with delayed rewards. Therefore, the REINFORCE algorithm [36] is introduced to compute the derivate of the object $J(\theta)$, *w.r.t* the parameters θ as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{p_\theta}(a_{1:N})[R(S) \sum_{i=1}^{N} \nabla_\theta log \pi_\theta(a_i|h_i)], \quad (15)$$

where a_i corresponds to actions $\boldsymbol{a}_{audio}^i$ and $\boldsymbol{a}_{visual}^i$. h_i corresponds to the hidden state representation $\hat{h}_{audio}^i$ and $\hat{h}_{visual}^i$. Therefore, the Eq.(11) can be denoted as follows:

$$\nabla_\theta J(\theta) = \mathbb{E}_{p_\theta}(a_a^{1:N}, a_v^{1:N})[R(S) \sum_{i=1}^{N} \nabla_\theta log \pi_\theta(a_a^i, a_v^i|h_a^i, h_v^i)]. \qquad (16)$$

Algorithm 1 RLLD algorithm
___
Input: The video sequential $\{A_t, V_t\}_{t=1}^{T}$, and the corresponding original category labels $y \in \{0, 1\}^C$, max epoch T_{max}

Parameter: The denoising networks parameters θ, the task network parameters W

Output: Network parameters θ and W for denoising and task modules
___
1: Let $t = 0$.
2: **while** not convergent or $t \leq T_{max}$ **do**
3:     Generate action $\boldsymbol{a}_{audio}^i$ and $\boldsymbol{a}_{visual}^i$ by the denoising network;
4:     Removing the noisy labels according to Eq.(5) and get the inter-reward according to Eq.(9);
5:     Train the AVVP network using the Eq.(8) to obtain W.
$$W = W + \alpha \nabla_W \mathcal{L}_{a,v}.$$
6:     Validate the AVVP model to obtain the reward $R(S)$ according to Eq.(11).
7:     Evaluate the gradients of the denoising network by Eq.(15),
$$\theta = \theta + \eta \nabla_\theta J(\theta).$$
8:     $t = t + 1$
9: **end while**
10: **return** The denoising model and the AVVP model.
___

Since Eq.(12) involves the expectation over high-dimensional action, which is hard to compute directly, a Monte Carlo approximation for the above quantity is:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta log \pi_\theta(a_a^i, a_v^i | h_a^i, h_v^i) R(S), \quad (17)$$

where N is the step length of one episode. Finally, we summarize our proposed RLLD approach for label denoising in AVVP in **Algorithm 1**.

3.5. Discussion

As a novel framework for AVVP, aimed at eliminating noisy labels in audio and visual modalities, RLLD stands out from previous methods. We initially integrate label denoising and AVVP into a unified framework. We employ reinforcement learning to explore effective label denoising strategies and leverage the feedback from AVVP to guide the learning of these strategies. Ultimately, there remains a crucial question regarding the motivation behind our proposed RLLD.

Why do we utilize reinforcement learning to learn the label denoising policy for AVVP? The primary reason is that the objective of label denoising in AVVP is to enhance video parsing performance. Reinforcement learning emphasizes goal-directed learning through interaction more than other machine learning paradigms [30]. Therefore, the improvement in parsing accuracy can be directly aligned with the goal (reward) in reinforcement learning. Secondly, in practical scenarios, we often lack knowledge about which labels are noisy in audio and visual modalities. The inherent characteristic of reinforcement learning is exploratory learning. Therefore, we employ reinforcement learning and design a soft inter-reward to discover strategies for automatically identifying noise labels specific to each modality. Finally, the label denoising module generates the denoising policy, and the validation results from the AVVP module guide the learning of the denoising policy. This architecture seamlessly integrates the denoising module and AVVP module into a joint learning framework, significantly simplifying modality-specific label denoising for AVVP.

4. Experiments

To evaluate the performance of RLLD for AVVP, we conduct comprehensive experiments. In this section, we present the details of experiment. All experiments are conducted using the PyTorch framework on two NVIDIA GeForce RTX 4090 GPUs under Pytorch platform.

4.1. Experimental Settings

Dataset and metrics. The *Look, Listen, and Parsing* (LLP) dataset [31] is utilized to evaluate our proposed method. This dataset is a weakly supervised audio-visual event dataset containing 11,849 videos across 25 categories. Diverse event labels are collected to annotate each video, including *human speaking, baby crying*, and *car running, etc*. Notably, over 7,000 videos contain multiple event labels. The LLP dataset partitions the entire dataset into training, validation, and testing subsets. The validation and testing subsets include both audio and visual events. The validation subset comprises 649 videos, while the testing dataset comprises 1,200 videos. In line with previous studies [31, 6, 40], we employ segment-level and event-level F-scores for all modalities as evaluation metrics. The mIoU threshold is set to 0.5 for computing event-level F-scores. The overall performance across all events is computed by considering the results of all audio and visual events, denoted as Event@Audio-Visual. Averaging the metrics for audio, visual, and audio-visual results provides the Type@Audio-Visual performance measure.

Implementation Details. Drawing parallels with prior works in AVVP, we utilize pre-trained ResNet512 [13] and 3D ResNet [33] to extract visual features, while pre-trained VGGish [14] is employed for audio feature extraction. The learning rate for the label denoising model is set to *1e-4*. The hyper-parameters of α_1 and α_2 are 0.6 and 0.4. The hyper-parameters of the task model follow the JoMoLD configuration. To address the state transitions issue in the label denoising module, we randomly sample 128 samples

as a batch and 1/4 of the samples are overlapped in the next state. For the audio-specific denoising branch, we utilize the mean F-score (*0.1) of segment-level under the audio type on the validation dataset as feedback to guide audio-specific label denoising. Similarly, the mean F-score (*0.1) of event-level under the visual type on the validation dataset is used to guide visual-specific label denoising for the visual branch. In instances where the sampled action removes all labels, we maintain the labels unchanged, providing a feedback reward of -1. We utilize Adam to optimize both the denoising and task models, ensuring their optimal performance.

Comparison Methods. In this study, we compare our proposed RLLD model with for state-of-the-art label denoising methods for AVVP: JoMoLD [6], MA [37] and with the enhanced contrastive learning (CL) loss. These methods aim to address modality-specific noise labels. The HAN utilizes label smoothing to address modality-specific challenges. JoMoLD and MA collaboratively exchange audio and visual modalities to identify noise labels. Additionally, we select weakly-supervised video parsing MM-Paramid [40], HAN [31] and AVSDN [21] as baselines for comparison. To further demonstrate the efficacy of our approach, we compare it with several state-of-the-art audio-visual event parsing methods: HAN with Co-teaching+ [41] and JoCoR [35] strategies, MGN [25], DHHN [18], CVCMS [22], CPSP [43], and poiBin [27]. The HAN utilizes label smoothing to address modality-specific challenges. JoMoLD and MA collaboratively exchange audio and visual modalities to identify noise labels. In our model, after removing modality-specific noise labels, the visual and audio-specific labels are directly utilized as ground truth for training the AVVP model. What needs to be pointed out is that our proposed module is a label denoising method. To validate the effectiveness of our method, we also conducted experiments by integrating our proposed label denoising method into two state-of-the-art (SOTA) methods, VALOR [42] and CPSP [43]. These two SOTA models were selected as the task network. We employed the denoising setting to train the models, demonstrating that our proposed label denoising method can further enhance the performance of the SOTA AVVP models.

4.2. Results

Table 2 presents the F-score results (%) for audio-visual video parsing in comparison with the state-of-the-art methods. In each table, the best performance is highlighted in boldface. As can be observed from the prediction results, our proposed RLLD method outperforms the baseline methods on F-score evaluation metrics. Specifically, RLLD significantly outperforms AVSDN and AVE on audio event type, achieving 61.5% under segment level and 53.4% under event-level. From the results, it is noteworthy that
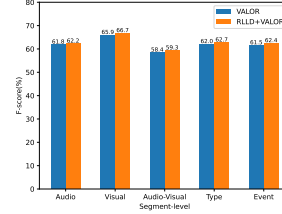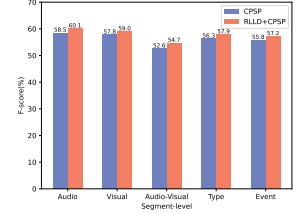

Figure 2. Segment-level results Figure 3. Segment-level results of VALOR and RLLD+VALOR. of CPSP and RLLD+CPSP.
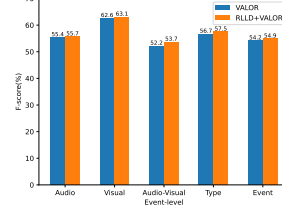

Figure 4. Event-level results of Figure 5. Event-level results of VALOR and RLLD+VALOR. CPSP and RLLD+CPSP.
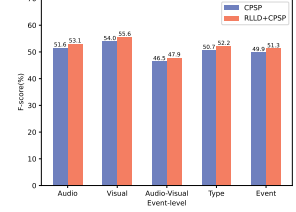
our approach performs marginally better than the state-of-the-art DHHN [18] on visual (64.6% *vs* 58.3% and 60.1% *vs* 55.1%) and on audio (61.6% *vs* 61.3% and 54.9% *vs* 54.0%). Our method outperforms HAN with two enhanced strategies (60.7% *vs* 57.7% and 53.4% *vs* 50.3%) on audio-visual event type. MA and JoMoLD are two label denoising methods designed to address the AVVP task. The performance of our method is also superior to the most comparable methods under the same setting (w/o contrastive learning (CL) loss). Based on the results, we can conclude that our method achieves overall state-of-the-art performance.

The superior performance of our proposed RLLD can be attributed to two primary reasons. Firstly, we formulate the problem of label denoising as an agent exploration decision-making process. This approach allows the agent to automatically learn the optimal policy for removing noise labels, rather than relying on direct comparison between audio and visual modalities. Secondly, we directly utilize the validation results as rewards to guide the learning of the label denoising policy. This design ensures that the proposed RLLD is directly optimized for enhancing parsing performance. The results clearly demonstrate that leveraging validation results is more effective in guiding the reinforcement learning agent to learn the optimal label denoising policy, ultimately leading to superior video parsing results.

4.3. Enhance the State-of-the-art Model

Our proposed method is a label denoising technique that can further enhance any AVVP models. To demonstrate the capabilities of our label denoising approach, we integrated it into two state-of-the-art (SOTA) models: VALOR [42] and CPSP [43].

For the 'RLLD+VALOR' setting, we maintain the pseudo-label extraction process from the pre-trained large

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
AVSDN	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MM-Paramid	60.9	54.4	50.0	55.1	57.6	52.7	51.8	44.4	49.9	50.5
HAN+Co-teaching+	59.4	56.7	52.0	56.0	56.3	50.7	53.9	46.6	50.4	48.7
HAN+JoCoR	61.0	58.2	53.1	57.4	57.7	52.8	54.7	46.7	51.4	50.3
MGN	60.8	55.4	50.4	55.5	57.2	51.1	52.4	44.4	49.3	49.1
CVCMS	59.2	59.9	53.4	57.5	58.1	51.3	55.5	46.2	51.0	49.7
DHHN	61.3	58.3	52.9	57.5	58.1	54.0	55.1	47.3	51.5	51.5
MA	59.8	57.5	52.6	56.6	56.6	52.1	57.4	45.8	50.8	49.4
JoMoLD	60.6	62.2	56.0	59.6	58.6	53.1	58.9	49.4	53.8	51.4
poiBin	63.1	63.5	57.7	61.4	60.6	54.1	60.3	51.5	55.2	52.3
MA+CL	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6
JoMoLD+CL	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5
RLLD (ours)	61.6	64.6	58.2	61.5	60.7	54.9	60.1	50.6	55.2	53.4
RLLD+CL (ours)	**63.4**	**65.1**	**58.7**	**62.4**	**61.9**	**55.8**	**61.2**	**51.2**	**56.1**	**54.0**
CPSP	58.5	57.8	52.6	56.3	55.8	51.6	54.0	46.5	50.7	49.9
CPSP + ours	**60.1**	**59.0**	**54.7**	**57.9**	**57.2**	**53.1**	**55.6**	**47.9**	**52.2**	**51.3**
VALOR	61.8	65.9	58.4	62.0	61.5	55.4	62.6	52.2	56.7	54.2
VALOR + ours	**62.2**	**66.7**	**59.3**	**62.7**	**62.4**	**55.7**	**63.1**	**53.7**	**57.5**	**54.9**

Table 1. Audio-visual video parsing F-score results (%) in comparison and the enhancing results with the state-of-the-art methods on the testing set of LLP.

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
RLLD w/o initialized labels	61.2	63.4	57.0	60.5	58.3	53.8	58.4	48.9	53.7	51.8
RLLD w/o soft inter-reward	61.8	64.2	57.5	61.1	60.1	54.6	59.3	50.1	54.6	53.2
Full Setting (RLLD+CL)	**63.4**	**65.1**	**58.7**	**62.4**	**61.9**	**55.8**	**61.2**	**51.2**	**56.1**	**54.0**

Table 2. Ablation results of the proposed modules on the LLP testing set.

model, eliminate the 'label elaboration' section in VALOR, and then, following Seq_LE [12], select the top-k labels as relevant based on their probability values, ensuring that the sum of the top-k probabilities does not exceed 0.5. In this configuration, RLLD replaces the 'label elaboration' section in VALOR. For the 'RLLD+CPSP' setting, we retain the PSP and PSA models from CPSP and utilize the denoising results of RLLD to train CPSP during the classification stage. The quantitative analysis results are shown in the last two parts of Table 1, and the visualization results are shown in Figure 2-5. We can conclude that our proposed label denoising method effectively enhances the performance of the existing AVVP models.

4.4. Ablation Study

The state representation of our proposed model incorporates both features and initialized labels. To assess the impact of each component within this representation, we conducted a rigorous ablation study by defining variations of RLLD. Specifically, we eliminated the initialized labels of RLLD. Specifically, we eliminated the initialized labels
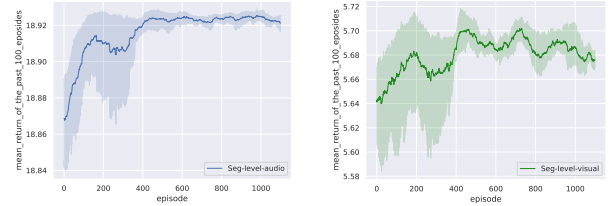

Figure 6. The audio curve under Figure 7. The visual curve under segment-level during the train-segment-level during the train-ing. ing.

in the RLLD variant 'w/o initialized labels'. We conducted this ablation study on the LLP dataset, evaluating performance using the F-score metric. The results, summarized in Table 2, reveal that the ablated versions exhibit inferior performance compared to the complete RLLD. The ablation study highlights the crucial role of the components within the state representation.

Additionally, to evaluate the impact of the soft inter-reward, we conducted an ablation study that excluded it,

Figure 8. The event@AV curve under segment-level during the training.

Figure 9. The event@AV curve under event-level during the training.
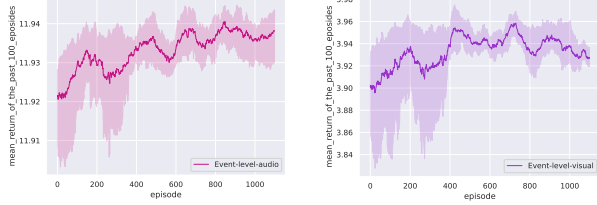

Figure 10. The audio curve under event-level during the training.

Figure 11. The visual curve under event-level during the training.

focusing only on the terminal reward. The results of this experiment are also presented in Table 2. We find that removing the soft inter-reward led to a relatively significant decrease in experimental performance. Therefore, we can draw the conclusion that the soft inter-reward contributes significantly to enhancing the label denoising performance.

4.5. Convergence Analysis

To further demonstrate the convergence of our proposed RLLD, we present convergence curves on the LLP dataset under various metrics. Specifically, we calculate the F-score for the past 100 episodes between the predicted results and the ground truth of the validation dataset. To expedite the training process and reduce computational costs, we randomly select one-third of the training set as a subset for one-epoch model training.

To obtain a representative convergence curve, we randomly initialize the model five times. The solid line represents the mean of these five experiments (0.1 smoothing), while the shaded area represents the corresponding standard deviation. The length of each episode is set to 32. The training process are shown in Figure 6-11. These figures reveal that the F-score for segment/Event audio metrics gradually increases with an increasing number of training episodes. For example, in the one-epoch validation feedback setting, the F-score of Segment-level@audio increases from 18.87 to 19.93. These convergence curves collectively demonstrate the convergence of our proposed method.

4.6. Limitation Discussion

Since there is no prior knowledge available to determine which labels are noise labels, a significant number of steps are required to explore and identify a strategy for noise label recognition. Consequently, the proposed RLLD may incur a longer training duration to achieve optimal performance. It is worth noting that the high time complexity associated with reinforcement learning remains an open problem. Therefore, further investigations are necessary to address this issue and enhance the efficiency of RLLD.

5. Conclusion

In this paper, we introduce a novel label denoising method for AVVP. We formulate the label denoising task as a sequential decision-making process and introduce a validation task to steer the learning of the denoising policy. Our approach bridges the gap between label denoising and AVVP, providing a unified framework that directly addresses parsing enhancements in AVVP. We evaluate our proposed method on LLP datasets and showcase its effectiveness. An ablation study further verifies the significance of initialized labels and the meticulously designed soft inter-reward in the state space. Furthermore, we conduct experiments to demonstrate that our proposed label denoising method can further bolster AVVP models. Future work will concentrate on exploring enhanced reward mechanisms to improve policy learning for label denoising.

Acknowledgements

References

[1] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *IEEE international conference on computer vision*, pages 609–617, 2017. 3

[2] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. *Neural Information Processing Systems*, pages 892–900, 2016. 3

[3] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021. 1

[4] D. A. Bulkin and J. M. Groh. Seeing sounds: visual and auditory interactions in the brain. *Current Opinion in Neurobiology*, 16(4):415–419, 2006. 1

[5] H. Chen, D. Zhu, G. Zhang, W. Shi, X. Zhang, and J. Li. Cm-cs: Cross-modal common-specific feature learning for audio-visual video parsing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023. 1

[6] H. Cheng, Z. Liu, H. Zhou, C. Qian, W. Wu, and L. Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 431–448. Springer, 2022. 2, 3, 4, 6, 7

[7] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *ACM International Conference on Multimedia*, pages 3884–3892, 2020. 3

[8] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 3

[9] Y. Fan, Y. Wu, Y. Lin, and B. Du. Revisit weakly-supervised audio-visual video parsing from the language perspective. In *Neural Information Processing Systems*, 2023. 3

[10] J. Gao, M. Chen, and C. Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18827–18836, 2023. 3

[11] R. Gao and K. Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15490–15500, 2021. 3

[12] Y. Gao, K. Wang, and X. Geng. Sequential label enhancement. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2022. 8

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *IEEE international conference on acoustics, speech and signal processing*, pages 131–135, 2017. 6

[15] A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall. Probabilistic future prediction for video scene understanding. In *European conference on computer vision*, pages 767–785, 2020. 1

[16] D. Hu, F. Nie, and X. Li. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 3

[17] R. A. Jacobs and C. Xu. Can multisensory training aid visual learning? a computational investigation. *Journal of Vision*, 19(11):1–1, 2019. 1

[18] X. Jiang, X. Xu, Z. Chen, J. Zhang, J. Song, F. Shen, H. Lu, and H. T. Shen. Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In *ACM International Conference on Multimedia*, pages 719–727, 2022. 1, 3, 4, 7

[19] S. Lee, J. Chung, Y. Yu, G. Kim, T. Breuel, G. Chechik, and Y. Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 3

[20] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 3

[21] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2002–2006, 2019. 7

[22] Y.-B. Lin, H.-Y. Tseng, H.-Y. Lee, Y.-Y. Lin, and M.-H. Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Neural Information Processing Systems*, pages 11449–11461, 2021. 3, 7

[23] M. Ling and X. Geng. Soft video parsing by label distribution learning. *Frontiers of Computer Science*, 13:302–317, 2019. 1

[24] J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 4133–4143, 2021. 1

[25] S. Mo and Y. Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. *Neural Information Processing Systems*, pages 34722–34733, 2022. 1, 3, 4, 7

[26] A. Nguyen, S. Järvelä, C. Rosé, H. Järvenoja, and J. Malmberg. Examining socially shared regulation and shared physiological arousal events with multimodal learning analytics. *British Journal of Educational Technology*, 54(1):293–312, 2023. 1

[27] K. K. Rachavarapu et al. Boosting positive segments for weakly-supervised audio-visual video parsing. In *IEEE/CVF International Conference on Computer Vision*, pages 10192–10202, 2023. 1, 2, 3, 4, 7

[28] T. Rahman, B. Xu, and L. Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *IEEE/CVF international conference on computer vision*, pages 8908–8917, 2019. 3

[29] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. 1

[30] R. Sutton and A. Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 9(5):1054–1054, 1998. 6

[31] Y. Tian, D. Li, and C. Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 436–454. Springer, 2020. 1, 3, 4, 6, 7

[32] F. Tosi, F. Aleotti, P. Z. Ramirez, M. Poggi, S. Salti, L. D. Stefano, and S. Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 4654–4665, 2020. 1

[33] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6

[34] Y. Wang and D. Wang. Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1381–1390, 2013. 3

[35] H. Wei, L. Feng, X. Chen, and B. An. Combating noisy labels by agreement: A joint training method with co-regularization. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 13726–13735, 2020. 7

[36] Williams and J. Ronald. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 3-4(8):229–256, 1992. 5

[37] Y. Wu and Y. Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 2, 3, 4, 7

[38] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[39] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 3

[40] J. Yu, Y. Cheng, R.-W. Zhao, R. Feng, and Y. Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *ACM International Conference on Multimedia*, pages 6241–6249, 2022. 1, 3, 6, 7

[41] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173, 2019. 7

[42] L. Yung-Hsuan, C. Yen-Chun, and F. W. Yu-Chiang. Modality-independent teachers meet weakly-supervised audio-visual event parser. In *Neural Information Processing Systems*, pages 1–19, 2023. 2, 4, 7

[43] J. Zhou, D. Guo, and M. Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7239–7257, 2023. 3, 7

[44] J. Zhou, D. Guo, Y. Zhong, and M. Wang. Improving audio-visual video parsing with pseudo visual labels. *arXiv preprint arXiv:2303.02344*, 2023. 1

[45] J. Zhou, D. Guo, Y. Zhong, and M. Wang. Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling. *International Journal of Computer Vision*, page Early Access, 2024. 3