Boosting Positive Segments for Weakly-Supervised Audio-Visual Video Parsing SUPPLEMENTARY MATERIAL

Kranthi Kumar Rachavarapu
Indian Institute of Technology Madras, India
kranthi.rachavarapu@gmail.com

Rajagopalan A. N.
Indian Institute of Technology Madras, India
raju@ee.iitm.ac.in

S1. Architecture of HAN baseline

In our work, we use the Hybrid Attention Network (HAN) architecture of Tian *et al*. [4] for weakly-supervised AVVP, as our goal is to improve the performance of a weakly-supervised model without necessarily relying on architectural changes. For the sake of completeness, we will now provide a detailed description of the architecture below.

Feature Extraction. Pre-trained audio CNN ($\mathbf{\Phi}_a$) and visual CNN ($\mathbf{\Phi}_v$) are employed to extract deep features for each segment. For any video, $f_t^a = \mathbf{\Phi}_a(A_t) \in \mathbb{R}^{d_a}$ and $f_t^v = \mathbf{\Phi}_v(V_t) \in \mathbb{R}^{d_v}$ are features of t-th audio and visual segments, respectively.

Feature Aggregation. To further capture cross modal information and inform the network about the most relevant temporal segments, we employ attentive feature fusion based on self-attention [6] to compute aggregated features as,

$$\hat{\mathbf{f}}_t^a = \mathbf{f}_t^a + \mathbf{\Phi}_{Att}(\mathbf{f}_t^a, \mathbf{F}^a, \mathbf{F}^a) + \mathbf{\Phi}_{Att}(\mathbf{f}_t^a, \mathbf{F}^v, \mathbf{F}^v) \tag{1}$$

$$\hat{\mathbf{f}}_t^v = \mathbf{f}_t^v + \mathbf{\Phi}_{Att}(\mathbf{f}_t^v, \mathbf{F}^v, \mathbf{F}^v) + \mathbf{\Phi}_{Att}(\mathbf{f}_t^v, \mathbf{F}^a, \mathbf{F}^a). \tag{2}$$

Here, $\mathbf{\Phi}_{Att}(.)$ is scalar-dot-product attention defined as,

$$\mathbf{\Phi}_{Att}(\mathbf{f}_q, \mathbf{F}_k, \mathbf{F}_v) = \text{Softmax}(\mathbf{f}_q \mathbf{F}_k^T / d)\mathbf{F}_v. \tag{3}$$

where $\mathbf{f}_q, \mathbf{F}_k, \mathbf{F}_v$ are d-dimensional key, query and value vectors, respectively.

Segment-level Event Prediction. Segment-level event probabilities are computed using a linear classifier with Sigmoid activation on aggregated features. The classifier transforms the input high-dimensional features into a C-dim vector which can be interpreted as logit vector indicating label distribution. Segment-level event probabilities are computed using Sigmoid operation on these final logits for each of the segment as,

$$\hat{\mathbf{p}}_t^m = \mathbf{\Phi}_c(\hat{\mathbf{f}}_t^m) \in \mathbb{R}^C, \quad t \in [1, T],\ m \in \{a, v\}, \tag{4}$$

where $\mathbf{\Phi}_c : \mathbb{R}^d \mapsto \mathbb{R}^C$ is a linear classifier.

Weakly-Supervised Event Prediction. As only video-level labels are available during training, we adopt attentive multi-modal Multi-Instance Learning (MMIL) to predict video-level event probabilities. First, modality-level labels are computed using attentive-pooling over temporal segments in each of the modality. Specifically, video-level event probabilities for audio, visual modalities of a video are computed as,

$$\hat{\mathbf{P}}^a = \sum_t w_t^a \hat{\mathbf{p}}_t^a \in \mathbb{R}^C\,, \qquad \hat{\mathbf{P}}^v = \sum_t w_t^v \hat{\mathbf{p}}_t^v \in \mathbb{R}^C \tag{5}$$

where $w_t^a, w_t^v \in \mathbb{R}^C$ are attention weights (over temporal segments) computed using a fully connected layer. Final video-level event probability is computed using attentive-pooling over modalities as $\hat{\mathbf{P}} = w_a \hat{\mathbf{P}}^a + w_v \hat{\mathbf{P}}^v$, where $w_a, w_v \in \mathbb{R}^c$ are attention weights over modality. We minimize the binary cross-entropy loss between predicted video-level event probability vector $\hat{\mathbf{P}}$ and weak video-level label $\mathbf{Y}$, given by,

$$\mathcal{L}_{\text{MIL}}^{\text{Att}} = CE(\hat{\mathbf{P}}, \mathbf{Y}) \tag{6}$$

S2. Modeling number of Positive segments as Poisson Binomial Distribution

For any given video, the segment level event probabilities follow Bernoulli distribution with the success probability of $\hat{\mathbf{p}}_t^m(c) \in [0, 1]\ \ \forall t \in [T],\ m \in \{a, v\}, c \in [C]$. Thus, the label distribution of all the segments of an event-c are independent

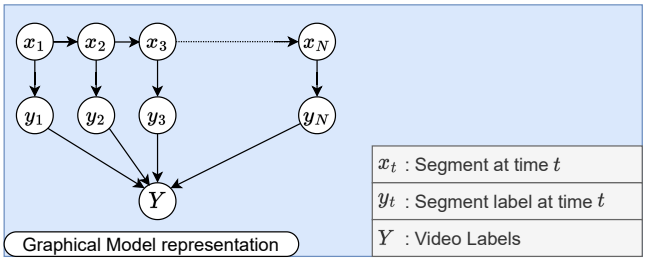

Figure S1: Graphical model representation of a video with N segments. Here, a segment label y_t is conditionally independent of other segments given the segment x_t.

and non-identical Bernoulli random variables, as each segment has a different success probability $\hat{\mathbf{p}}_t^m(c)$ as shown in Fig. S1. Without loss of generality, We describe the modeling for a particular event class c and ignore the variable c for the rest of the discussion.

Let z be a random variable (RV) denoting the number of *positive* segments with the event in a video *i.e.* $z = \sum_{\forall t,m} \hat{\mathbf{y}}_t^m$, where $\hat{\mathbf{y}}_t^m \sim \text{Bernoulli}(\hat{\mathbf{p}}_t^m)$ indicates whether t-th segment in modality-m has the event or not. The RV z follows the Poisson Binomial distribution, which can be computed exactly from the segment probabilities as described below. The distribution of z can be obtained using the characteristic functions of random variables $\hat{\mathbf{y}}_t^m$ as follows. The characteristic function of $\hat{\mathbf{y}}_t^m$ is, $\varphi_{\hat{\mathbf{y}}_t^m}(n) = \mathbb{E}[e^{in\hat{\mathbf{y}}_t^m}] = 1 - \hat{\mathbf{p}}_t^m + \hat{\mathbf{p}}_t^m e^{in}$, where $i = \sqrt{-1}$. The characteristic function of z is,

$$\varphi_z(n) = \mathbb{E}[e^{inz}] = \sum_{k=0}^{N} \mathbf{P}_z(k; N) e^{ink} \tag{7}$$

where $\mathbf{P}_z(k; N)$ is the probability of exactly k positive segments and $N - k$ negative segments in a video with N segments. Now, The characteristic function of $z = \sum_{\forall t,m} \hat{\mathbf{y}}_t^m$ can be alternatively expressed as,

$$\varphi_z(n) = \mathbb{E}(e^{in(\sum_{\forall t,m} \hat{\mathbf{y}}_t^m)}) \tag{8}$$

$$= \prod_{t,m} \varphi_{\hat{\mathbf{y}}_t^m}(n) = \prod_{t,m} (1 - \hat{\mathbf{p}}_t^m + \hat{\mathbf{p}}_t^m e^{in}). \tag{9}$$

We know (7) equals to (9). Therefore, we obtain,

$$\sum_{k=0}^{N} \mathbf{P}_z(k; N) e^{ink} = \prod_{\forall t,m} (1 - \hat{\mathbf{p}}_t^m + \hat{\mathbf{p}}_t^m e^{in}). \tag{10}$$

By substituting $n = \omega l$, $\omega = \frac{2\pi}{N+1}$ in (10), we obtain,

$$\sum_{k=0}^{N} \mathbf{P}_z(k; N) e^{i\omega lk} = \prod_{\forall t,m} (1 - \hat{\mathbf{p}}_t^m + \hat{\mathbf{p}}_t^m e^{i\omega l}). \tag{11}$$

In (11), the left side is the DFT of the sequence $P_z(k; N)$, which indicates the probability of k-positive segments out of N segments. Therefore, we can apply IDFT on both sides to recover $P_z(k; N)$ as,

$$\mathbf{P}_z(k; N) = \frac{1}{N+1} \sum_{l=0}^{N} e^{-i\omega lk} \left[\prod_{\forall t,m} (1 - \hat{\mathbf{p}}_t^m + \hat{\mathbf{p}}_t^m e^{i\omega l}) \right]. \tag{12}$$

S3. Temporal Action Localization

S3.1. Problem Formulation

Temporal Action Localization (TAL) aims to localize the start and end timestamps of action instances and recognize their categories simultaneously in untrimmed videos. In this paper, we consider weakly-supervised TAL that aims to localize and classify all action instances in a video given only video-level category labels during training.

Specifically, given a video $\mathbf{V}$ of T non-overlapping temporal segments $\{x_t)\}_{t=1}^{T}$ of visual frames, our objective is to classify each frame into one of the C possible events. Thus during evaluation, we need to identify *instance-level* event labels, $\mathbf{y}_t = \{0, 1\}^C$, for t-th frame. In weakly-supervised TAL, for each video $\mathbf{V}$, we only have access to the corresponding

video-level event label $\mathbf{Y} = [Y_0, Y_1, \ldots, Y_C] \in \{0,1\}^C$ s.t $\sum_{c=1}^{C} Y_c = 1$, where $Y_c = 1$ if any of the segments in the video contains c-th event, otherwise $Y_c = 0$. These *weak* video-level labels only indicate whether an event occurred in the given video or not. During the evaluation, our model should predict the temporal location of activity instances, i.e., for a testing video, it outputs a set of tuples (t_s, t_e, ψ, c) where t_s and t_e are the start and end frames of action, c is the action label, and ψ is the activity score.

S3.2. Architecture

We adopt an attentive fusion based architecture to predict per-frame probabilities. Following the previous works on action recognition, For each video, we first divide it into non-overlapping segments to extract segment-level features. We extract segment-level features for both the RGB and flow streams using pre-trained networks as,

$$f_t^{RGB} = \mathbf{\Phi}^{RGB}(x_t) \in \mathbb{R}^{d'} \tag{13}$$

$$f_t^{Flow} = \mathbf{\Phi}^{Flow}(x_t) \in \mathbb{R}^{d''} \tag{14}$$

where $\Phi^{RGB}(.)$ and Φ^{Flow} are pre-trained video and optical flow feature extractors, respectively. We first extract higher-level features by processing each of these modalities independently as,

$$\hat{f}_t^{RGB} = \mathbf{\Phi}_0^{RGB}(f_t^{RGB}) \in \mathbb{R}^d \tag{15}$$

$$\hat{f}_t^{Flow} = \mathbf{\Phi}_0^{Flow}(f_t^{Flow}) \in \mathbb{R}^d \tag{16}$$

We then fuse cross-modal information using the attention framework as,

$$\hat{\mathbf{f}}_t^{RGB} = \hat{f}_t^{RGB} + \mathbf{\Phi}_{Att}^{RGB}(\hat{f}_t^{RGB}, \hat{f}_t^{Flow}) \tag{17}$$

$$\hat{\mathbf{f}}_t^{Flow} = \hat{f}_t^{Flow} + \mathbf{\Phi}_{Att}^{Flow}(\hat{f}_t^{Flow}, \hat{f}_t^{RGB}) \tag{18}$$

where, $\mathbf{\Phi}_{Att}^{RGB}$ and $\mathbf{\Phi}_{Att}^{Flow}$ are cross-modal attention network to fuse features across modality. To get frame-level action probabilities, we then use a linear classifier on each modality as,

$$\hat{\mathbf{p}}_t^m = \mathbf{\Phi}_c(\hat{\mathbf{f}}_t^m) \in \mathbb{R}^C, \quad t \in [1, T],\ m \in \{RGB, Flow\}, \tag{19}$$

where $\mathbf{\Phi}_c : \mathbb{R}^d \mapsto \mathbb{R}^C$ is a linear classifier with Softmax activation. The final frame-level probability is the average of probabilities across RGB and FLow modalities. As only video-level labels are available during training, we adopt our proposed Poisson-binomial based (MMIL) along with attentive-MIL to predict video-level event probabilities.

S3.3. Modeling for Poisson binomial based MIL formulation for Multi-class classification

The proposed Poisson Binomial based MIL formulation (§3.4 in the Main paper) is for binary classification. But other action recognition problems [1, 5] are multi-class classification problems. In such cases, we can instead use the Poisson multinomial distribution (PMD) [2]. However, computing PMD requires enumerating all possible outcomes and quickly becomes impractical for even moderately sized problems. For example, in the Audio-Visual Event Localization (AVEL) task, where each video consists of 10 segments and 28 event categories, the total possible outcomes are 348 million, which makes it infeasible to use PMD in practice. Therefore, we need alternative approaches for modeling multi-class distributions even in moderate-scale problems.

To address this limitation, we propose a simple alternative of using the Poisson Binomial distribution for modeling multi-class classification problems. Given that in a multi-class classification problem, only one event occurs at any given time, we treat all the rest of the $C - 1$ events as a single background class. This enables us to use Poisson Binomial based modeling efficiently for multi-class classification problems. Note that we are given the foreground event label during training, making this formulation feasible. This approach simplifies the modeling process while maintaining accuracy, making it suitable for even large-scale multi-class classification problems, as shown in our experiments on the Temporal Action Localization task.

S3.4. Dataset

We evaluate our approach on THUMOS14 [1]. This dataset contains 200 validation videos for training and 213 testing videos for testing with 20 action categories. This is a challenging dataset with around 15.5 activity segments and 71% background activity per video. For feature extraction, we sample the video into non-overlapping 16 frame segments for both the RGB and the flow stream. Following the previous works, we use the I3D network pre-trained on the Kinetics dataset to extract both RGB and flow features. During training, we randomly sample 500 snippets, and during evaluation, we take all the snippets. We use the Adam optimizer with learning rate 0.0001.

S4. Qualitative comparison of Audio-Visual Video Parsing

We provide additional audio-visual video parsing results in Fig. S2, Fig. S3, Fig. S4, Fig. S5 on multiple videos. These examples also contain a few cases where our method fails to detect and localize events accurately.

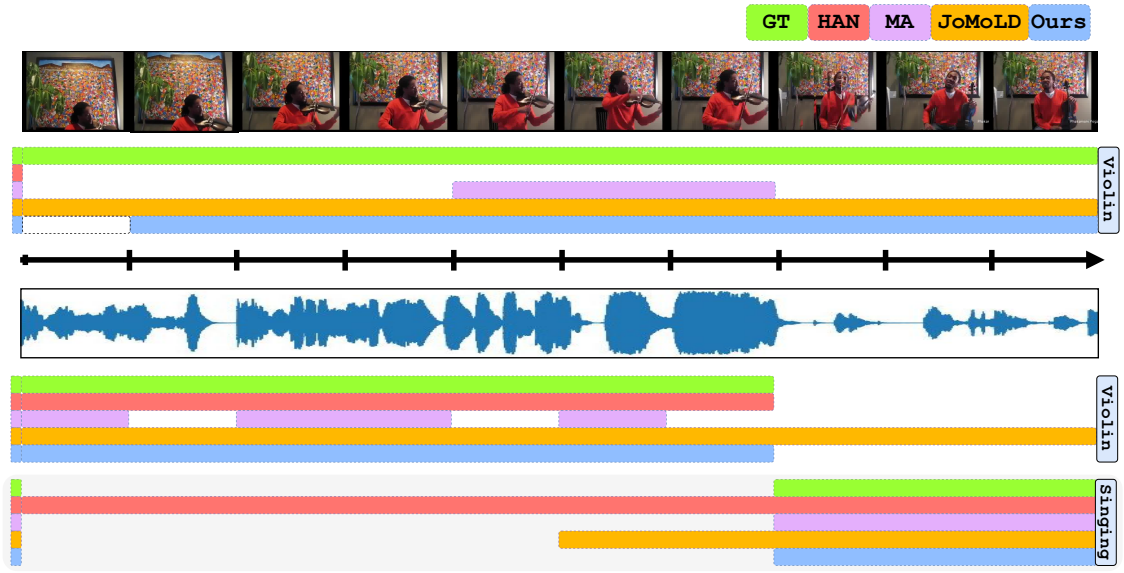

Figure S2: Audio-Visual Video Parsing results of our methods and HAN [4], MA [7], JoMoLD [3] on a video with *"Violin"*(Visual, Audio), *"Singing"* (Audio) events.

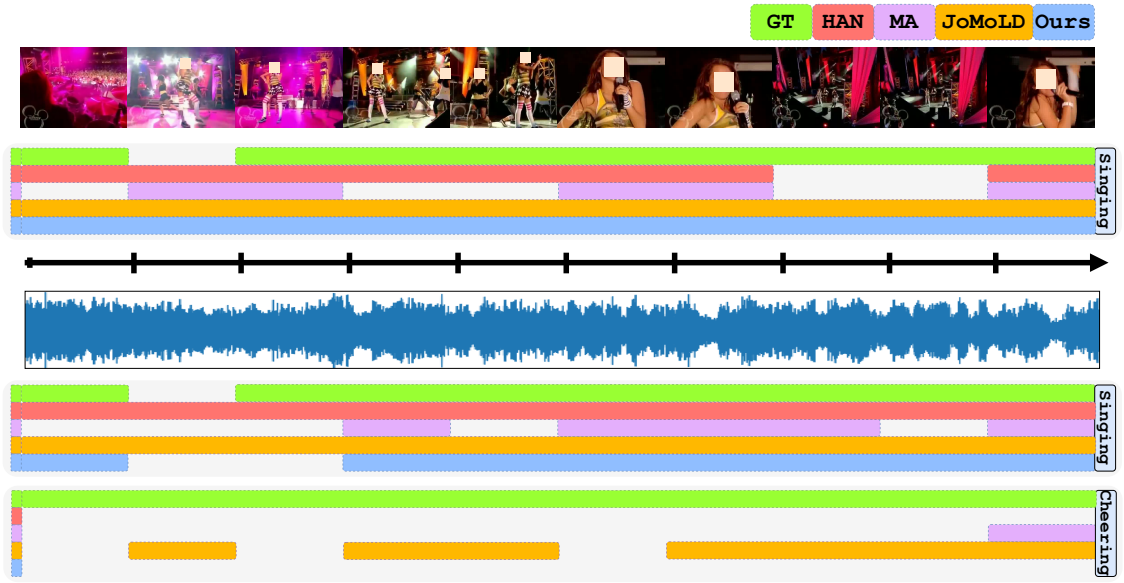

Figure S3: Audio-Visual Video Parsing results of our methods and HAN [4], MA [7], JoMoLD [3] on a video with *"Singing"*(Visual, Audio), *"Cheering"* (Audio) events.

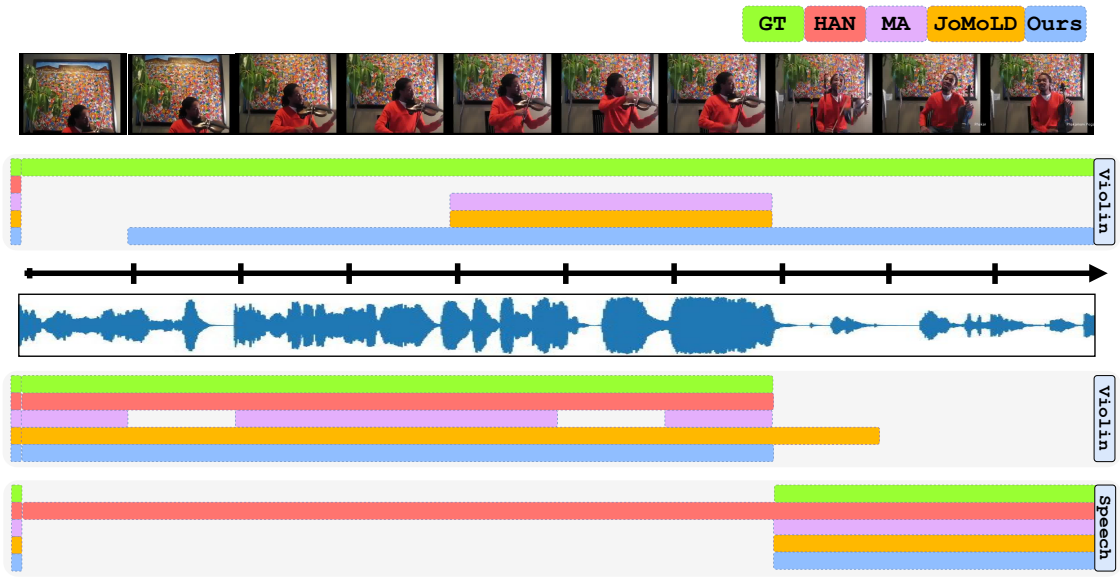

Figure S4: Audio-Visual Video Parsing results of our methods and HAN [4], MA [7], JoMoLD [3] on a video with *"Violin"* (Audio, Visual) and *"Speech"* (Audio) events.

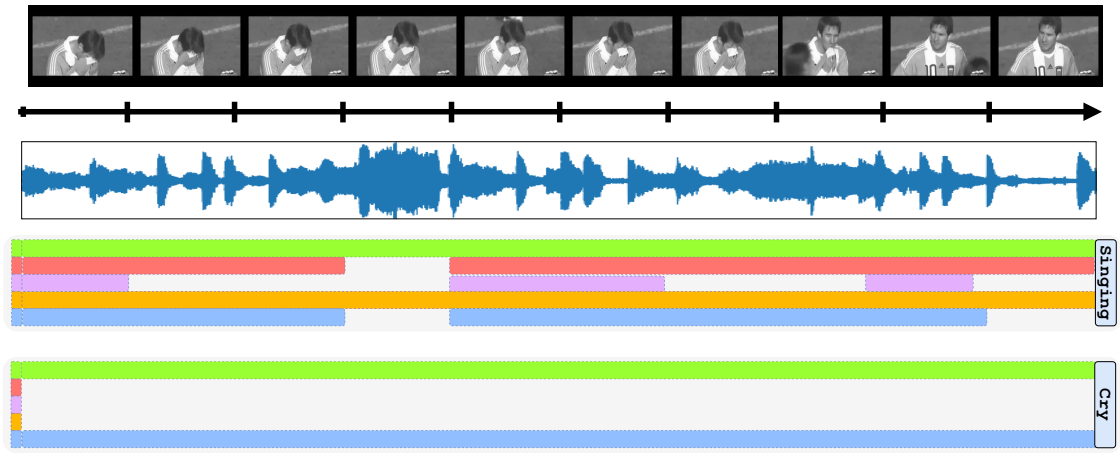

Figure S5: Audio-Visual Video Parsing results of our methods and HAN [4], MA [7], JoMoLD [3] a video with *"Singing"* (Audio) and *"Cry"* (Audio) events.

S5. Analysis of Failure Cases and Future Directions

Expanding on §6 from the main paper, we show the confusion matrices on selected events on segment predictions in Figure S6. Musical events are confused with one another, and usually, the models favor only a subset of events when more than one event is present in the segment. This is a direct result of assuming and modeling that the events are independent. AVVP has a strong label correlation. For instance, Singing/Speech occurs almost always with other events such as commentary (basketball, motorcycle), and pets (cat/dog) or musical instruments. Our method estimates τ^* that is always a local optimum, which may limit the performance. Thus adopting prior-based generative modeling appears to be an exciting future direction.

References

[1] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 3

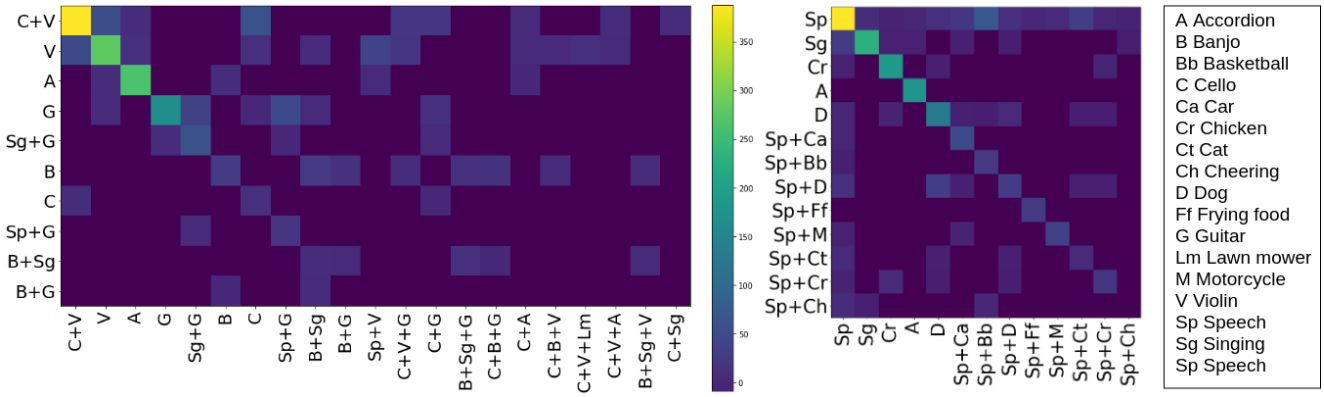

Figure S6: Analysis of failure cases.

[2] Zhengzhi Lin, Yueyao Wang, and Yili Hong. The poisson multinomial distribution and its applications in voting theory, ecological inference, and machine learning. *arXiv preprint arXiv:2201.04237*, 2022. 3

[3] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Advances in Neural Information Processing Systems*, 2022. 4, 5

[4] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer, 2020. 1, 4, 5

[5] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 3

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[7] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 4, 5