Supplementary Materials
Joint-Modal Label Denoising for Weakly-Supervised Audio-Visual Video Parsing

Haoyue Cheng[1,2] Zhaoyang Liu[2] Hang Zhou[3] Chen Qian[2]
Wayne Wu[2] Limin Wang[1,4]✉

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] SenseTime Research [3] CUHK - Sensetime Joint Lab [4] Shanghai AI Laboratory
chenghaoyue98@gmail.com zyliumy@gmail.com zhouhang@link.cuhk.edu.hk
qianchen@sensetime.com wuwenyan0503@gmail.com lmwang@nju.edu.cn

The appendix provides more visualizations and analyses to show deep insights into our method. Sec. 1 exhausts the details of optimizing the network after performing modality-specific label denoising in each training iteration. In Sec. 2, we conduct more ablation studies for our method. To further compare JoMoLD with other methods (*i.e.*, HAN [1] and MA [3]), Sec. 3 provides more specific visualization cases.

1 Details of optimizing network $\mathcal{F}$

This section elaborates on the details of optimizing parsing network $\mathcal{F}$.

Data Input. As described in Sec. 3.1 of the main paper, for a given video, pretrained off-the-shelf networks extract segment-level audio and visual features $\mathbf{f}^a = \{f_1^a, ..., f_t^a, ..., f_T^v\}$, $\mathbf{f}^v = \{f_1^v, ..., f_t^v, ..., f_T^v\}$, where t denotes the segment timestamp and T represents the total number of segments. The fixed local features are fed into the network $\mathcal{F}$.

Feature Aggregation. Previous work [1] proves the significance of aggregating temporal context and leveraging the clues in different modalities. We define a function *Attn* to represent the widely used attention mechanism:

$$Attn(q, \mathbf{K}, \mathbf{V}) = Softmax(\frac{q\mathbf{K}^T}{d})\mathbf{V}, \tag{1}$$

where d represents the dimension of vector q. Local features f_t^a and f_t^v are then enhanced by the following way:

$$\begin{aligned} \hat{f}_t^a &= f_t^a + Attn(f_t^a, \mathbf{f}^a, \mathbf{f}^a) + Attn(f_t^a, \mathbf{f}^v, \mathbf{f}^v), \\ \hat{f}_t^v &= f_t^v + Attn(f_t^v, \mathbf{f}^v, \mathbf{f}^v) + Attn(f_t^v, \mathbf{f}^a, \mathbf{f}^a), \end{aligned} \tag{2}$$

The enhanced features $\hat{f}_t^a, \hat{f}_t^v$ are context-aware and have better capabilities of identifying the events occurred at t-segment.

Model Output. The audio-visual video parsing task predicts the event categories in each segment. In network $\mathcal{F}$, a shared fully-connected layer projects

✉: Corresponding author.

enhanced audio and visual features $\hat{f}_t^a$, $\hat{f}_t^v$ to label space. As a multi-label multi-class classification task, the sigmoid function is applied to further output the probabilities (between 0-1) for all event categories for each segment. We express this process as the following equation:

$$p_t^a = Sigmoid(FC(\hat{f}_t^a)), \quad p_t^v = Sigmoid(FC(\hat{f}_t^v)), \tag{3}$$

where $p_t^a, p_t^v \in (0,1)^C$. During training, since only video-level labels are available, parsing network $\mathcal{F}$ adopts an attentive MMIL Pooling mechanism to obtain audio-level, visual-level and video-level event probability $p^a, p^v, p \in (0,1)^C$ by gathering weighted average of segment-level event probabilities:

$$\begin{aligned} p^a[c] = \sum_{t=1}^{T} W_t^a[c]\, p_t^a[c], \quad p^v[c] = \sum_{t=1}^{T} W_t^v[c]\, p_t^v[c], \\ p[c] = \sum_{t=1}^{T} W_t^{av}[0,c] W_t^a[c]\, p_t^a[c] + W_t^{av}[1,c] W_t^v[c]\, p_t^v[c], \end{aligned} \tag{4}$$

where $W_t^a, W_t^v \in (0,1)^C$ and $W_t^{av} \in (0,1)^{2\times C}$ are temporal and audio-visual attention weights respectively. $W^a = \{W_t^a\}_{t=1}^T, W^v = \{W_t^v\}_{t=1}^T \in (0,1)^{T\times C}$ are derived from applying learnable MLPs on $\hat{f}_t^a, \hat{f}_t^v$, and normalized by softmax function on temporal axis. And $W^{av} = \{W_t^{av}\}_{t=1}^T \in (0,1)^{T\times 2\times C}$ are derived from applying another learnable MLP layer on features $\hat{f}_t^a, \hat{f}_t^v$, then normalized on modality axis.

Training Losses. We optimize the network in a batch manner. Let B denote the batch size, and $\mathbf{P}^a$, $\mathbf{P}^v$, $\mathbf{P} \in (0,1)^{B\times C}$ represent the audio-level, visual-level and video-level event probabilities of a batch samples. The labels $\mathbf{Y}^a$, $\mathbf{Y}^v \in \{0,1\}^{B\times C}$ are the refined audio-level and visual-level labels obtained by modality-specific label denoising. $\mathbf{Y} \in \{0,1\}^{B\times C}$ represent the original video-level labels. We can then optimize network $\mathcal{F}$ with audio-level loss $\mathcal{L}_a$, visual-level loss $\mathcal{L}_v$, and video-level loss $\mathcal{L}_s$ using binary cross-entropy loss:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_a + \mathcal{L}_v + \mathcal{L}_s \\ &= -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} \mathbf{Y}^a[b,c] log(\mathbf{P}^a[b,c]) \\ &\quad - \frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} \mathbf{Y}^v[b,c] log(\mathbf{P}^v[b,c]) \\ &\quad - \frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} \mathbf{Y}[b,c] log(\mathbf{P}[b,c]). \end{aligned} \tag{5}$$

2 More Ablation Studies

In this section, we explore more ablation studies to demonstrate the rationality of our method, which are not displayed in the main paper due to space limitation. Segment-level metrics are reported.

Study the Impact of Cross-modal Attention during ***Calculating Forward Loss***. As stated in Sec. 3.3 of the main paper, we skip cross-modal attention when calculating forward loss in modality-specific label denoising. Cross-modal attention interferes with the event predictions of two modalities, and produces inaccurate modality-specific losses and further inaccurate noisy labels. Results in Table 2.1 quantify the effectiveness of removing cross-modal attention.

Study the Generality of JoMoLD Equipped with Other Baselines. We mainly utilizes HAN [1] as the backbone since it's a widely used baseline. To further validate the generality, we change different backbones with JoMoLD. We modify two models from similar tasks as backbones, *i.e.*, 1) AVE [2]: audio-visual event localization task; 2) and AVSlowFast [4]: audio-visual action recognition task. Table 2.2 confirms that our approach works well for different baselines.

Study the Impact of Different Batch Sizes. We study the impact of different batch sizes on the final results in Table 2.3. The results of the models trained on smaller sizes are slightly lower than that on batch size 128, which is the optimal setting in our experiments. Two reasons can explain this: 1) Noises might not be uniformly distributed in a smaller batch; 2) There are more round-off errors for smaller batch sizes when determining the number of noises in a batch. The results of model trained on a larger batch size fluctuate within acceptable ranges.

Table 2.1: **Study the effectiveness of skipping cross-modal attention.**

Forward Loss for Label Denoising	Audio	Visual	Audio-Visual	Type@AV	Event@AV
Not Skip cross-modal attention	60.3	60.0	55.1	58.9	57.9
***Skip* cross-modal attention**	**61.3**	**63.8**	**57.2**	**60.8**	**59.9**

Table 2.2: **Study the generality of JoMoLD on other backbones.**

Method	Audio	Visual	Audio-Visual	Type@AV	Event@AV
AVE [2]	49.9	37.3	37.0	41.4	43.6
AVE + JoMoLD	**50.8**	**39.5**	**39.8**	**43.4**	**45.9**
AVSlowFast [4]	47.2	50.8	39.8	45.9	47.0
AVSlowFast + JoMoLD	**48.9**	**60.1**	**43.7**	**50.9**	**52.1**

Table 2.3: **Study the impact of different batch sizes.**

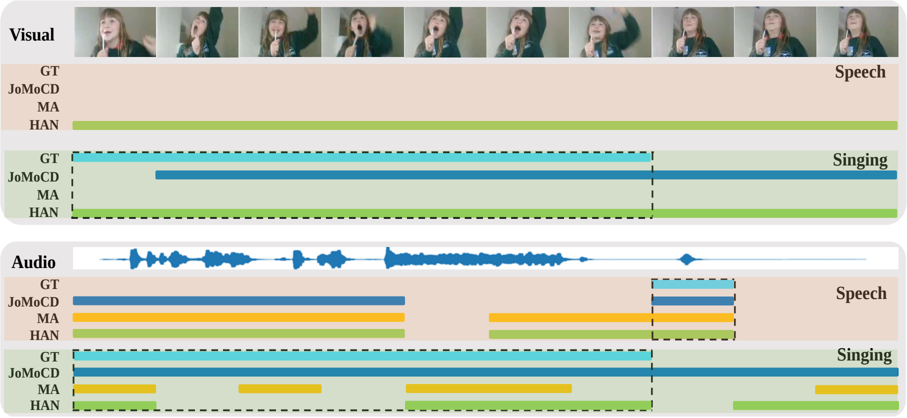
Batch size	Audio	Visual	Audio-Visual	Type@AV	Event@AV
32	61.3	63.1	56.4	60.3	59.6
64	61.4	63.2	57.0	60.5	59.7
128	**61.3**	**63.8**	**57.2**	**60.8**	**59.9**
256	61.4	63.6	57.1	60.8	59.5

3 Additional Qualitative Analyses

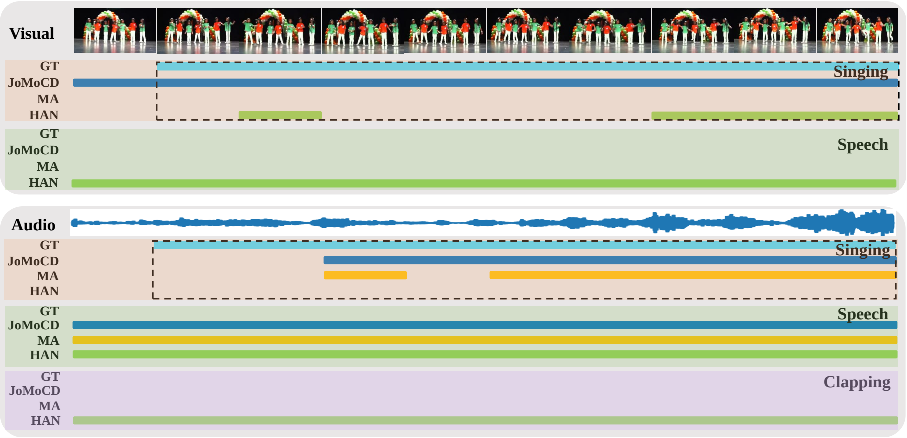
In this section, we present additional visualization cases to compare our JoMoLD with other methods on video parsing and label denoising.

3.1 Visualizations of Video Parsing

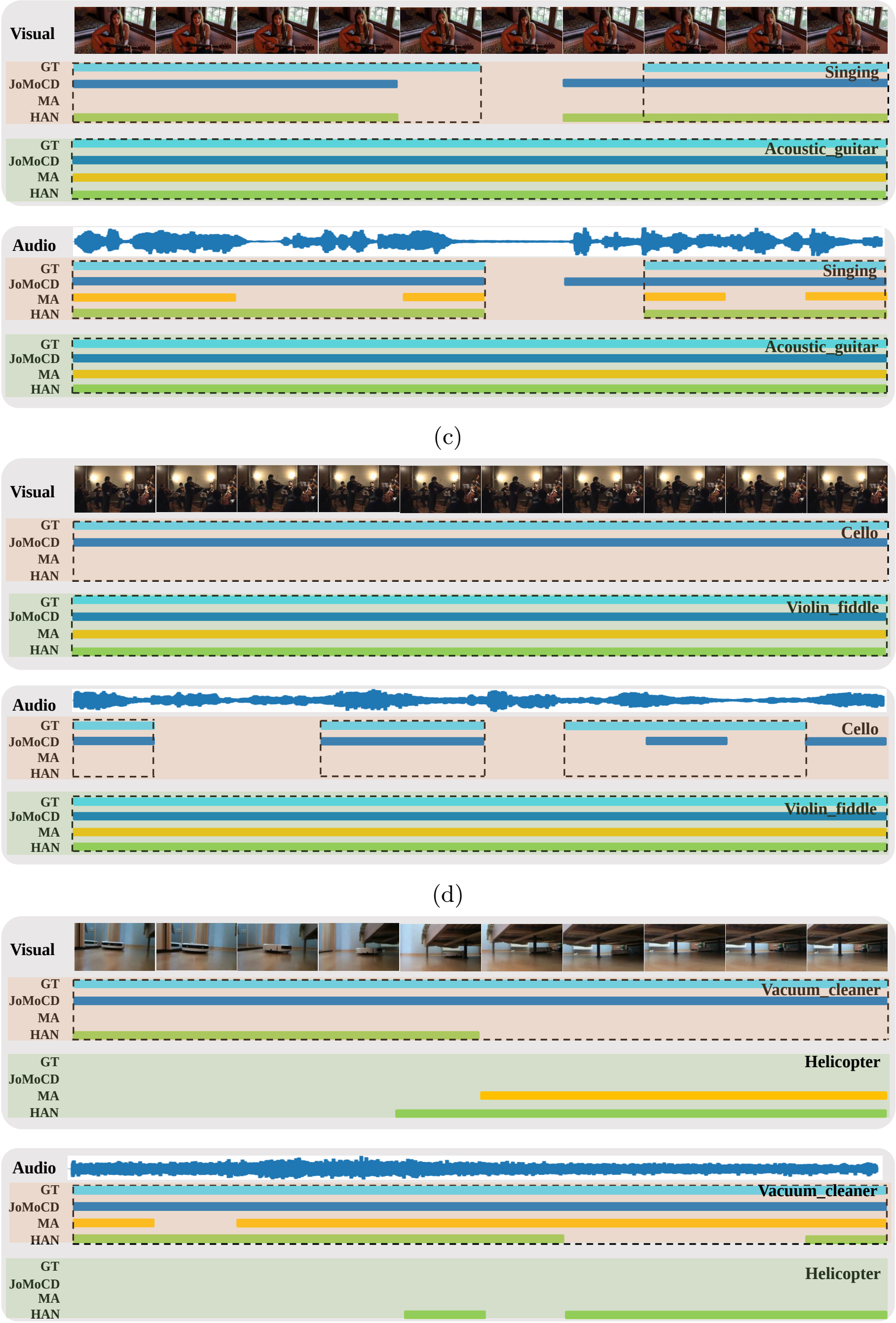
We visualize the video parsing results of JoMoLD, HAN [1] and MA [3] on different examples. "GT" denotes the ground truth annotations. Each video lasts for 10 seconds. Our method achieves more accurate parsing performance by acquiring reliable modality-specific supervision during training.



(a)



(b)



(c)

(d)

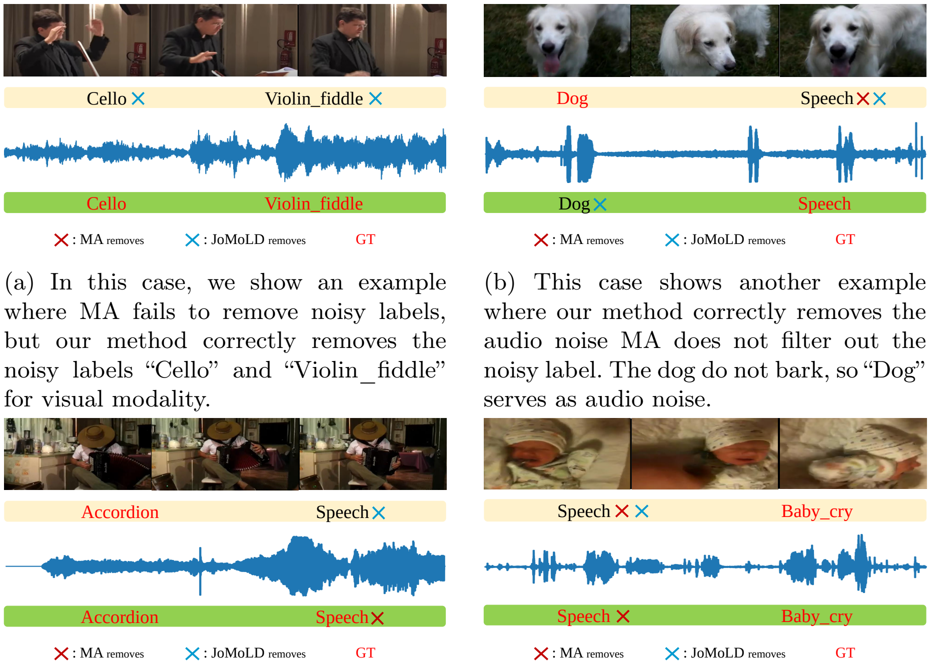
(e)

Fig. 3.1: **We visually compare JoMoLD with HAN, MA and ground truths.** In some easier examples, such as 3.1d and 3.1e, we achieve superior detection performance. In some difficult cases, the overall performance of JoMoLD is still better than HAN and MA.

3.2 Visualization of Label Denoising

In this section, we show that JoMoLD is superior to MA [3] in most cases when it comes to determining modality-specific noisy labels.

On the one hand, as MA keeps cross-modal attention when performing modality-specific label denoising, the two modalities interfere with each other to cause inaccurate denoising results. While JoMoLD avoids cross-modal interference. On the other hand, MA adopts the naively trained baseline to determine the noisy labels. It trains the baseline with original videos but exchanges the audio tracks of two unrelated videos during label denoising, which leads to the gap between training and denoising. In contrast, there is no gap for JoMoLD, which consistently processes the original videos when training and denoising. Meanwhile, JoMoLD adopts a dynamic manner to analyze the loss patterns of two modalities and remove noisy labels, which has a higher tolerance for denoising errors.



(a) In this case, we show an example where MA fails to remove noisy labels, but our method correctly removes the noisy labels "Cello" and "Violin_fiddle" for visual modality.

(b) This case shows another example where our method correctly removes the audio noise MA does not filter out the noisy label. The dog do not bark, so "Dog" serves as audio noise.

(c) In the case, MA mistakenly removes the correct label but remains the noisy label. The person in the picture doesn't speak and another person is speaking off-screen. So "Speech" is a visual noise.

(d) This case presents an example that our method correctly identifies the noisy modality but MA treats them both as noise. A parent is speaking off-screen so "Speech" is a visual noise.

Fig. 3.2: **Label denoising comparison between MA and JoMoLD.** We list four cases to illustrate different kinds of mistakes made by MA and avoided by our JoMoLD.

References

1. Tian, Y., Li, D., Xu, C.: 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. pp. 436–454. Springer (2020) 1, 3, 4
2. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 247–263 (2018) 3
3. Wu, Y., Yang, Y.: Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1326–1335 (2021) 1, 4, 6
4. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020) 3