Weakly-Supervised Audio-Visual Video Parsing with Prototype-based Pseudo-Labeling
SUPPLEMENTARY MATERIAL

1. Model

Our model for weakly-supervised AVVP closely follows the architecture of Tian et al. [14] but with necessary modifications as described below.

Feature Extraction. Akin to previous approaches [14, 17], 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 Projection. These features capture generic information as they are extracted using pre-trained models. To extract task-specific features, we employ projection head networks to refine them. These audio feature sequence ($\mathbf{F}^a$) and visual feature sequence ($\mathbf{F}^v$) of a video are defined as

$$\mathbf{F}^a = \{\hat{\mathbf{f}}_t^a\}_{t=1}^T \ , \ \mathbf{F}^v = \{\hat{\mathbf{f}}_t^v\}_{t=1}^T, \tag{1}$$

where $\hat{\mathbf{f}}_t^a = \mathbf{\Phi}_a^{\text{proj}}(f_t^a) \in \mathbb{R}^d$, $\hat{\mathbf{f}}_t^v = \mathbf{\Phi}_v^{\text{proj}}(f_t^v) \in \mathbb{R}^d$ are refined audio,visual features and $\mathbf{\Phi}_a^{\text{proj}} : \mathbb{R}^{d_a} \mapsto \mathbb{R}^d$, $\mathbf{\Phi}_v^{\text{proj}} : \mathbb{R}^{d_v} \mapsto \mathbb{R}^d$ are audio, visual projection networks, respectively. Using this, audio and visual features are projected into a common d-dimensional embedding space.

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 [15] inspired from [14, 17].These aggregation features are computed as,

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

$$\mathbf{f}_t^v = \hat{\mathbf{f}}_t^v + \mathbf{\Phi}_{Att}(\hat{\mathbf{f}}_t^v, \mathbf{F}^v, \mathbf{F}^v) + \mathbf{\Phi}_{Att}(\hat{\mathbf{f}}_t^v, \mathbf{F}^a, \mathbf{F}^a). \tag{3}$$

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

$$\mathbf{\Phi}_{Att}(\mathbf{f}_q, \mathbf{F}_k, \mathbf{F}_v) = \text{Softmax}(\mathbf{f}_q \mathbf{F}_k^T / d)\mathbf{F}_v. \tag{4}$$

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

Weakly-Supervised Event Prediction. Segment-level event probabilities are computed using a linear classifier with Sigmoid activation on aggregated features as,

$$\hat{\mathbf{p}}_t^m = \mathbf{\Phi}_c(\mathbf{f}_t^m) \in \mathbb{R}^C, \quad t \in [1, T], \ m \in \{a, v\}, \tag{5}$$

where $\mathbf{\Phi}_c : \mathbb{R}^d \mapsto \mathbb{R}^C$ is a linear classifier. As only video-level labels are available during training, we adopt attentive multimodal Multi-Instance Learning (MMIL) to predict video-level event probabilities. First, modality-level labels are computed using attentive pooling over temporal segments in each modality. Specifically, video-level event probabilities for audio and visual modalities of a video are computed as

$$\hat{\mathbf{P}}^a = \sum_t w_t^a \hat{\mathbf{p}}_t^a \in \mathbb{R}^C \ , \qquad \hat{\mathbf{P}}^v = \sum_t w_t^v \hat{\mathbf{p}}_t^v \in \mathbb{R}^C \tag{6}$$

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 $\mathbf{W}$, given by,

$$\mathcal{L}_{\text{MIL}} = CE(\hat{\mathbf{P}}, \mathbf{W}). \tag{7}$$

Momentum encoder follows the same structure and contains the above-described Feature Projection, Feature Aggregation, and Segment Classifier blocks. The weights of the momentum encoder are updated using an exponential moving average ensemble of MIL-based models from different training steps instead of using backpropagation.

2. Experimental Results

Comparison with off-the-shelf Pseudo-Labeling methods: Here, we perform experiments to analyze how effective the existing pseudo-labeling strategies are for the AVVP task. We compare our method with EM-MIL [9] and Poibin [12]. The results reported in Tab. 2 indicate that both EM-MIL and Poibin perform poorly on AVVP.

These methods [9, 12] are designed by incorporating strong priors specific to each WS task. E.g., EM-MIL implicitly assumes that all events co-occur (Eq.5 in [9]), which is not valid for AVVP as event occurrence is asynchronous. And Poibin [12] proportion of positives as the

Table 1. Architecture summary. $\mathbf{d}_{in}, \mathbf{d}_{out}$ stand input and output dimensions of feature maps, respectively. All layers use *Leaky-ReLU* activation.

Task	Name	Type	$\mathbf{d}_{in}$	$\mathbf{d}_{out}$
Feature Extraction	Audio ($\mathbf{\Phi}_a$)	Pretrained VGGish		
	Visual ($\mathbf{\Phi}_v$)	Pretrained ResNet-18		
Feature Refinement	Audio ($\mathbf{\Phi}_a^{\text{proj}}$)	Linear	128	256
		LayerNorm	-	256
		Linear	256	512
		LayerNorm	-	512
	Visual ($\mathbf{\Phi}_v^{\text{proj}}$)	Linear	2048	1024
		LayerNorm	-	1024
		Linear	1024	512
		LayerNorm	-	512
Feature Aggregation	Attentive Fusion ($\mathbf{\Phi}_{\text{Att}}$)	Multi-Head attention with 1-head		
Weakly Supervised Classification	Segment Classifier ($\mathbf{\Phi}_{\text{c}}$)	Linear+Sigmoid	512	25
	Attention-Temporal	Linear+Softmax	512	25
	Attention-Modality	Linear+Softmax	512	25

Table 2. Comparison with off-the-shelf Pseudo-Labeling methods.

Method	Audio		Visual		Audio-Visual		Type@AV		Event@AV	
	Seg.	Event	Seg.	Event	Seg.	Event	Seg.	Event	Seg.	Event
EM-MIL [9]	59.3	50.5	53.6	49.9	49.7	43.3	54.2	47.9	55.2	48.1
PoiBin [12]	63.1	54.1	63.5	60.3	57.7	51.5	61.4	55.2	60.6	52.3
Ours	**65.9**	**57.3**	**66.7**	**64.3**	**61.9**	**54.3**	**64.8**	**59.9**	**63.7**	**57.9**

Table 3. Effect of temperature τ on model performance.

τ	Audio		Visual		Audio-Visual		Type@AV		Event@AV	
	Seg.	Event	Seg.	Event	Seg.	Event	Seg.	Event	Seg.	Event
0.05	62.7	56.9	63.9	60.8	59.7	53.9	62.1	57.3	60.5	55.5
0.1	**65.9**	**57.3**	**66.7**	**64.3**	**61.9**	**54.3**	**64.8**	**59.9**	**63.7**	**57.9**
0.2	63.1	57.3	64.1	60.5	59.9	54.3	62.3	57.0	60.8	54.7
0.3	58.3	51.9	61.2	56.5	57.6	52.2	59.0	53.6	56.4	49.9

pseudo-labels, which does not provide stronger constraints on the segment-level labels. Also, PL methods, in general, may not help, as the traditional approach of directly using model predictions as ground truth was *not* effective (Tab.2 of the main paper). Thus, it is not PL per se that works- *how* these labels are obtained and utilized is paramount.

Analysis of τ. A smaller temperature τ in Eq. 10 of the main paper gives a more concentrated distribution, while a larger one makes it more uniform. We experiment with different values while generating soft pseudo labels in Tab. 3. Up to a point, performance improves on decreasing τ and is best for $\tau = 0.1$. We also find that soft labeling with a small temperature $\tau = 0.05$ performs similarly to hard labeling (see $\texttt{Base+PPL}_{\texttt{H}}$, Tab. 2 of main paper), which supports intuition.

Effect of Momentum Encoder. Features extracted from the momentum encoder are used for prototype feature generation, as described in Sec. 4 of the main paper. We experiment with and without a momentum encoder, and the results are reported in Table 4. Here, we can observe that a momentum-encoder-based setup improves performance.

Table 4. Effect of momentum encoder.

EMA	Segment-Level					Event-Level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
W/out	64.7	66.5	60.1	63.8	63.1	56.8	63.9	52.6	58.2	56.4
With	65.9	66.7	61.9	64.8	63.7	57.3	64.3	54.3	59.9	57.9

Failure cases: Expanding on limitations from Sec.7, we show two failure cases in Figure 1 (highlighted in red) below (zoom in for a better view). As our method relies on feature similarity for pseudo labeling, here it resulted in one false positive (in frame-2 of both videos) as it is very similar in appearance to its neighboring frame.

3. Results on Weakly-Supervised Temporal Action Localization task

We experiment on other weakly-supervised event localization tasks to validate the efficacy of the proposed approach. Here, we conduct a set of preliminary experiments on the Weakly-Supervised Temporal Action Localization (TAL) task. We perform preliminary experiments on the Temporal Action Localization (TAL) that aims to localize the start and end timestamps of action instances and recognize their categories simultaneously in untrimmed videos. Here, the input consists of two modalities - RGB and flow. RGB frames are sensitive to the scene content, whereas flow modality is appearance invariant and more sensitive to the motion. We experiment with the THUMOS14 dataset [5], which consists of videos with 100's of frames belonging to 20 ac-

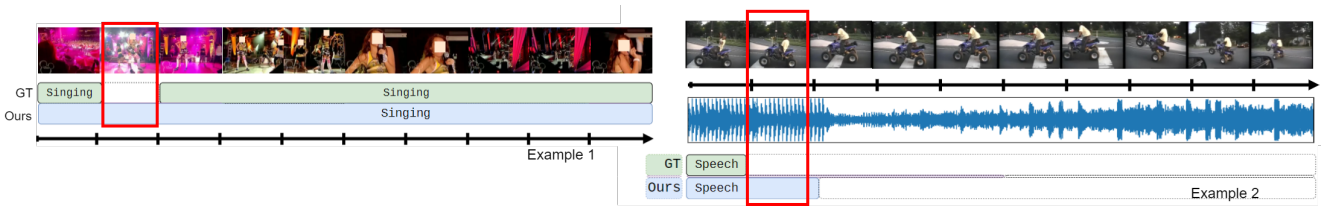

Figure 1. Failure cases of our method on two videos. Zoom in for a better view.

Table 5. Results on Temporal Action Localization task on THUMOS14 dataset [5]. The best and second-best results are shown in **bold** and underline, respectively.

Method	IoU							AVG		
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	[0.1:0.5]	[0.3:0.7]	[0.1:0.7]
Wang*et al.* [16]	44.4	37.7	28.2	21.1	13.7	-	-	20.6	-	-
W-TALC [11]	55.2	49.6	40.1	31.1	22.8	-	7.6	39.8	-	-
EM-MIL [9]	59.1	52.7	45.5	36.8	30.5	22.7	16.4	44.9	30.4	37.7
Nguyen et al.[10]	60.4	56	46.6	37.5	26.8	17.6	9	45.5	27.5	36.3
HAM-Net [6]	65.4	59	50.3	41.1	31	20.7	11.1	49.4	30.9	39.8
FTCL [2]	69.6	63.4	55.2	45.2	35.6	23.7	12.2	53.8	34.4	43.6
UGCT [18]	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6	43.6
DCC [7]	69.0	63.8	55.9	45.9	35.7	24.3	13.7	54.1	35.1	44.0
DGCNN [13]	66.3	59.9	52.3	43.2	32.8	22.1	13.1	50.9	32.7	41.3
Li et al. [8]	69.7	64.5	**58.1**	**49.9**	39.6	**27.3**	14.2	56.3	37.8	46.1
Huang et al. [4]	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
ASM-Loc [3]	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
DELU [1]	71.5	66.2	56.5	47.7	**40.5**	27.2	15.3	56.5	37.4	46.4
nPP (Ours)	**72.7**	**66.9**	57.9	46.9	37.4	26.8	**20.1**	56.4	**37.8**	**46.9**

tion categories. The video length varies significantly from a few seconds to minutes. The duration of an action instance also has a large variance, from shorter than one second to several minutes. We adopt the same architecture design as HAN [14] for this task.

We report the results for this setup in Table 5. We evaluate in terms of mean Average Precision (mAP) with different temporal Intersection over Union (tIoU) thresholds, which is denoted as mAP@α where α is the threshold. Our model, trained with our proposed approach from Sec. 4 of the main paper, achieves better or comparable performance than the current state-of-the-art model DELU [1] for tIOU of $0.1, 0.2, 0.3, 0.4$. Our model also shows more significant improvements at high threshold metrics tIoU=0.7, which implies that our action proposals are more complete.

References

[1] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 192–208. Springer, 2022. 3

[2] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20009, 2022. 3

[3] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13925–13935, 2022. 3

[4] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2022. 3

[5] 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. 2, 3

[6] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1637–1645, 2021. 3

[7] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19914–19924, 2022. 3

[8] Ziqiang Li, Yongxin Ge, Jiaruo Yu, and Zhongming Chen. Forcing the whole video as background: An adversarial learning strategy for weakly temporal action localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5371–5379, 2022. 3

[9] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 729–745. Springer, 2020. 1, 2, 3

[10] Phuc Xuan Nguyen, Deva Ramanan, and Charless C Fowlkes. Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5502–5511, 2019. 3

[11] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 3

[12] Kranthi Kumar Rachavarapu and A N Rajagopalan. Boosting positive segments for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10192–10202, 2023. 1, 2

[13] Haichao Shi, Xiao-Yu Zhang, Changsheng Li, Lixing Gong, Yong Li, and Yongjun Bao. Dynamic graph modeling for weakly-supervised temporal action localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3820–3828, 2022. 3

[14] 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, 3

[15] 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

[16] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017. 3

[17] 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. 1

[18] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 53–63, 2021. 3