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

Supplementary Material

We begin this supplementary document by expounding on the limitations of our proposed UWAV method. In the section that follows, we elaborate on the implementation details, compute environment used for implementation, and the training and inference times of our proposed method. Then, we quantitatively compare with VALOR on the LLP datset using better backbone features. In Section 6, we put forward studies showcasing the sensitivity of our method to the choice of the hyper-parameters α, W, followed by ablation studies about the various design choices of our model. Finally, we end this document by providing some qualitative visualizations of the predictions obtained by our method versus competing baselines on both the LLP and the AVE datasets.

The following summarizes the supplementary materials we provide:

- Limitations.
- Implementation details of UWAV.
- Details of our compute environment.
- Compute time analysis.
- Quantitative comparison using better backbone features.
- Studies on the sensitivity of UWAV to the choice of α, W.
- The Scalability of UWAV.
- Ablation studies on the different design choices.
- Qualitative results of UWAV versus competing methods for the AVVP task.

1. Limitations

Although UWAV achieves state-of-the-art results on the AVVP task, compared to competing methods, it requires additional training data to pre-train the pseudo-label generation module (on which we train for about 80 epochs).

2. Implementation Details of UWAV

To assess the effectiveness of our method, in line with prior work [11], each 10-second video in both the LLP [11] and AVE [10] datasets is split into 10 segments of one second each, where each segment contains 8 frames. The visual feature backbone for the LLP dataset is based on the ResNet-152 [3] network (pre-trained on the ImageNet dataset [1]) for extracting 2D-appearance features, and the R(2+1)D [12] network (pre-trained on the Kinetics-400 dataset [5]) for extracting features that capture the visual dynamics, respectively. The VGGish [4] network, pre-trained on the AudioSet dataset [2], is used to extract features from the audio, sampled at 16KHz. For the AVE dataset however, akin to prior work [6], we extract visual features from pre-trained CLIP [8] and R(2+1)D, while CLAP [13] is used to embed the audio stream. For both datasets, we set the number of encoder blocks L of the transformers in the pseudo-label generation module to 5, α for the Beta distribution in the feature mixup strategy to 1.7, and W in the class-balanced loss re-weighting step to 0.5. Both the pseudo-label generation modules and the inference modules are trained with the AdamW optimizer [7]. To train the model, we employ a learning rate scheduling strategy that warms up the learning rate for the first 10 epochs to its peak of $1e{-}4$ and then decays according to a cosine annealing schedule, to the minimum, which is set to $1e{-}5$ for the pseudo-label generation models and $5e{-}6$ for the inference model. We clip the gradient norm at 1.0 during training. For the LLP dataset, the training batch size is set to 64 and the total number of training epochs to 80 for both models, while the same is set to 16 and 80 for the AVE dataset.

Table A1. **Compute time analysis on the LLP dataset.** "Inference Time" denotes the time to evaluate all testing data.

Method	Training Time per Epoch	Inference Time
CoLeaf [9]	25 sec	24 sec
UWAV (Ours)	24 sec	20 sec

3. Details of Compute Environment

Our model is trained on a desktop computer with an Intel Core i7 CPU, with 32GB RAM, and a single NVIDIA RTX 3090 GPU.

4. Analysis of Compute Time

For a more holistic understanding of the performance of our method, we compare its training and inference times with the most recently published approach for the AVVP task, viz. CoLeaf [9] on the LLP dataset [11]. The results of this study are shown in Table A1. We see that our method's runtime performances are comparable with those of competing approaches, with notable inference time gains over the CoLeaf method [9].

5. Quantitative Comparison Using Better Backbone Features

We also quantitatively compare our proposed approach with VALOR on the LLP dataset using better backbone features, *i.e*. CLIP and CLAP as visual and audio feature backbones. As shown in Table A2, UWAV outperforms VALOR with 2.7 F-score improvement in segment-level Type@AV and

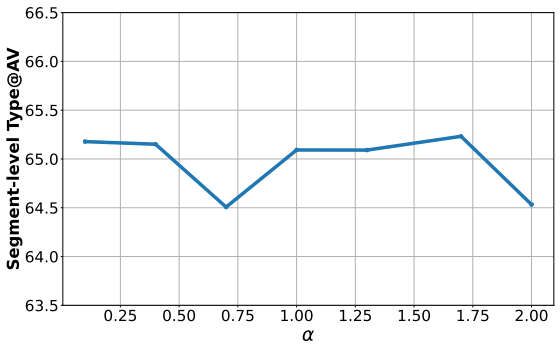


Figure A1. **Sensitivity of α in the uncertainty-weighted feature mixup on the LLP dataset.**

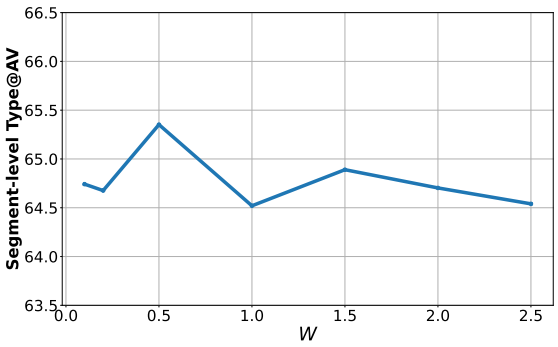


Figure A2. **Sensitivity of W in the class-balanced re-weighting on the LLP dataset.**

Table A2. **Comparison with VALOR on the LLP dataset.** † denotes using CLIP and CLAP features as input to the HAN model.

Method	Segment-level					Event-level				
	A	**V**	**AV**	**Type**	**Event**	**A**	**V**	**AV**	**Type**	**Event**
VALOR† [17]	68.1	68.4	61.9	66.2	66.8	61.2	64.7	55.5	60.4	59.0
UWAV† (Ours)	**68.9**	**72.3**	**65.6**	**68.9**	**68.3**	**63.5**	**68.7**	**59.6**	**63.9**	**62.4**

3.5 F-score improvement in event-level Type@AV.

6. Sensitivity to the Choice of α and W

To gain a better understanding of the effect of the choice of hyper-parameters on our model's performance, we evaluate the sensitivity of our model to the choice α in the uncertainty-weighted feature mixup and W in the class-balanced loss re-weighting strategy. When α is adjusted, class-balanced loss re-weighting is not applied. As shown in Figure A1, for the LLP dataset, when α increases from 0.1 to 2.0, segment-level Type@AV F-score first decreases to 64.5, then rises to a peak of 65.2 at $\alpha = 1.7$, and subsequently declines back to 64.5. On the other hand, Figure A2 illustrates the effect of varying W on the segment-level Type@AV F-score. The F-score reaches its maximum value of 65.3 when $W = 0.5$ and decreases as W becomes larger. When W is adjusted, the uncertainty-weighted feature mixup is not applied. These observations point towards the robustness of our model to the precise choice of these hyper-parameters. We observe similar trends for the AVE dataset as well. Hence, for best results, we select $\alpha = 1.7$ and $W = 0.5$ in all our experiments for both datasets.

7. The Scalability of UWAV

To evaluate the scalability of UWAV, we train the inference model (HAN) with less training data (Table A3b) as well as fewer event classes (Table A3a) on the LLP dataset by removing the training videos or event classes randomly. Even with only 60% of the training data, UWAV exhibits competitive performance. Moreover, UWAV shows a consistent performance lead against VALOR [17], irrespective of the number of event classes, with no change in training strategy or the core model structure.

Table A3. **The scalability of UWAV.**

(a) Training with different amounts of data.

Training Data Ratio	Segment-level				
	A	**V**	**AV**	**Type**	**Event**
100%	**64.2**	**70.0**	**63.4**	**65.9**	**63.9**
80%	63.4	69.2	62.5	65.0	63.0
60%	63.4	68.6	62.4	64.8	62.8

(b) Training with different number of classes.

	Segment-level Type F-score		
Number of Classes	25 (all events)	20	15
VALOR [17]	62.0	65.9	66.6
UWAV(Ours)	65.9	71.4	68.4

8. Ablation Studies

Ablation Study on All Metrics: In Table A4, we report the ablation study on all metrics for a more complete understanding. Coupled with our proposed class-balanced re-

Table A4. **Ablation study reported on all metrics.** "Binary" denotes training with binary pseudo-labels. "Soft" denotes training with uncertainty-weighted pseudo-labels.

Binary	Soft	Re-weight	Mixup	Segment-level					Event-level				
				A	**V**	**AV**	**Type**	**Event**	**A**	**V**	**AV**	**Type**	**Event**
✓				62.7	67.7	61.9	64.2	62.2	56.9	64.9	56.6	59.5	55.8
	✓			63.0	68.3	61.8	64.4	62.8	56.9	65.2	55.9	59.3	56.1
	✓	✓		63.6	69.5	63.0	65.4	63.1	57.9	66.4	57.0	60.4	56.9
	✓		✓	63.9	69.0	62.8	65.2	63.4	57.7	65.6	56.3	59.9	56.8
	✓	✓	✓	**64.2**	**70.0**	**63.4**	**65.9**	**63.9**	**58.6**	**66.7**	**57.5**	**60.9**	**57.4**

Table A5. **Ablation study of uncertainty-weighted mixup in Eq. 14 and Eq. 15. on the AVE dataset.**

Method	$\hat{p}_t^a, \hat{p}_t^v$	$\lceil \hat{p}_t^a \rceil, \lceil \hat{p}_t^v \rceil$
Acc.(%)	80.3	**80.6**

weighting strategy, the HAN model improves from 59.3 to 60.4 for the event-level Type@AV. On the other hand, by introducing the proposed uncertainty-aware mixup strategy, the event-level Type@AV increases from 59.3 to 60.

Ablation Study of the Uncertainty-weighted Mixup on the AVE Dataset: As shown in Table A5, our experiments reveal that using $\lceil \hat{p}_t^a \rceil$ and $\lceil \hat{p}_t^v \rceil$ as the segment-level pseudo labels instead of $\hat{p}_t^a$ and $\hat{p}_t^v$ for the uncertainty-weighted feature mixup strategy, in Eq. 14 and Eq. 15. in the main paper, results in a slightly better performance on the AVE dataset.

9. Qualitative Results

Figures A3, A4 show event predictions of our method versus competing baselines on sample videos from the LLP dataset [11]. Figure A5 shows the same, for sample videos on the AVE dataset [10]. As is evident from the figures, we see consistently accurate event-label predictions across different videos, while also generally accurately localizing them, the same is not the case for the baseline approaches. This feature is particularly prominent for instance, for the visual event classes in the first video in Figure A4, or the audio-visual events in the second video example in Figure A5. However, there remain challenging scenarios where almost all methods struggle, such as the audio events in the first video example in Figure A4, which we hope to address going forward.

Figure A3. **Comparison between predictions by UWAV and competing AVVP methods on the LLP dataset.** "GT": ground truth.

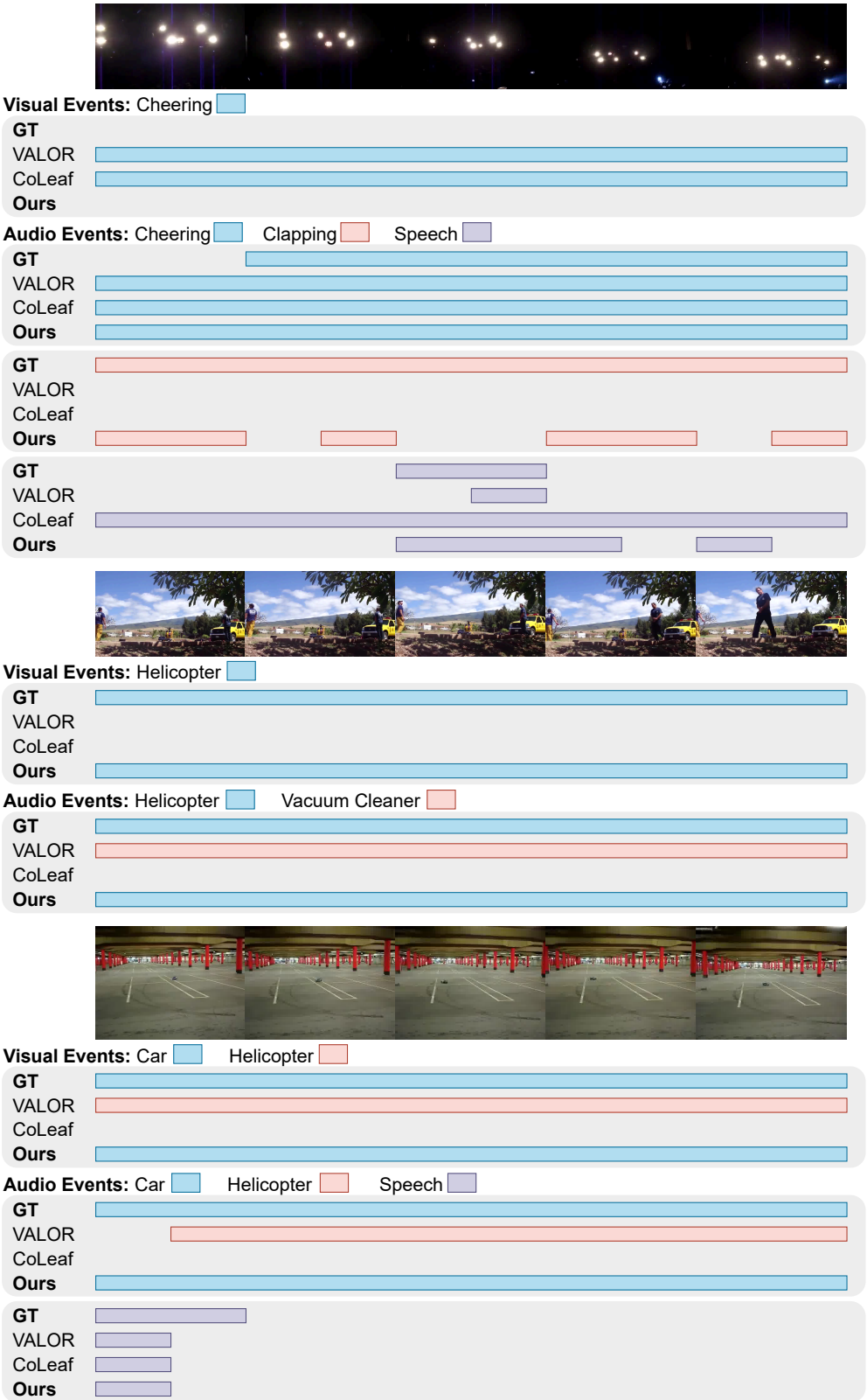
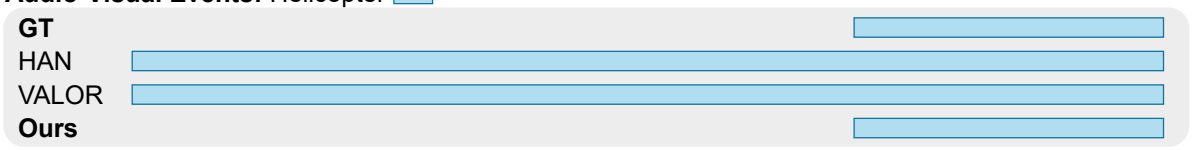
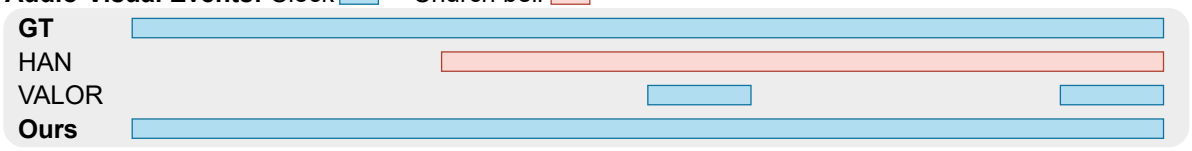


Figure A4. **Comparison between predictions by UWAV and competing AVVP methods on the LLP dataset.** "GT": ground truth.

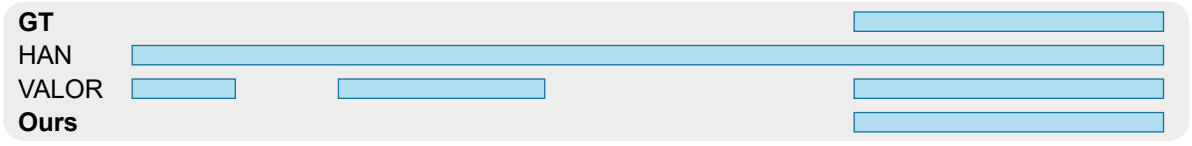
Audio-Visual Events: Helicopter

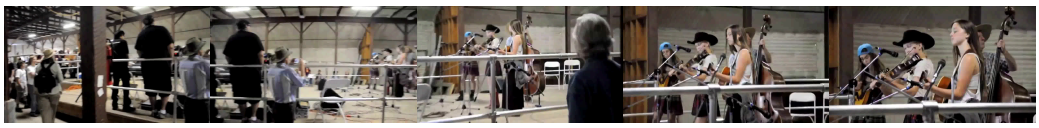


Audio-Visual Events: Clock Church bell

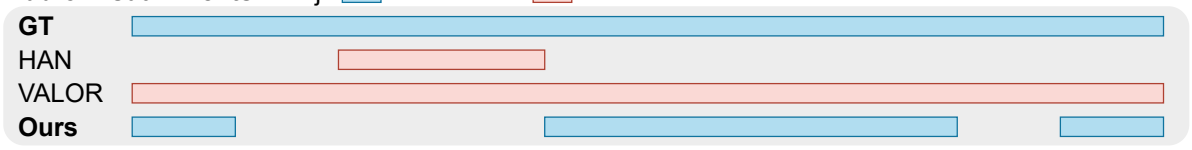


Audio-Visual Events: Church bell





Audio-Visual Events: Banjo Mandolin



Audio-Visual Events: Bus

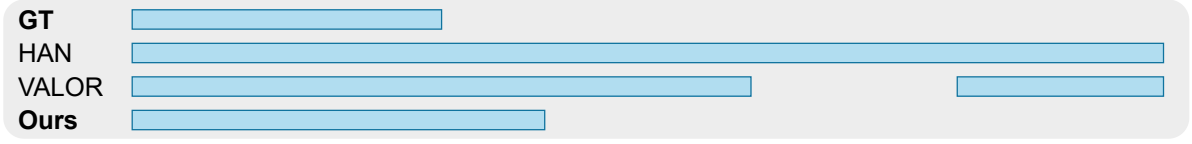


Figure A5. **Qualitative comparison between predictions by UWAV and previous methods on the AVE dataset.** "GT": ground truth.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017. 1

[5] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[6] Yung-Hsuan Lai, Yen-Chun Chen, and Frank Wang. Modality-independent teachers meet weakly-supervised audio-visual event parser. In *NeurIPS*, 2023. 1

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[9] Faegheh Sardari, Armin Mustafa, Philip JB Jackson, and Adrian Hilton. Coleaf: A contrastive-collaborative learning framework for weakly supervised audio-visual video parsing. In *ECCV*, 2024. 1

[10] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1, 3

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

[12] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 1

[13] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023. 1