In this supplementary material, we present additional experimental results, including a parameter study of the λ used in our semantic-aware optimization strategy (Eq. 12 in our main paper) and more ablation studies of the proposed LEAP block. Furthermore, we analyze the computational complexity of the model. At last, we provide more qualitative examples and analyses of audio-visual video parsing to better demonstrate the superiority and interpretability of our method.

A Parameter study of λ

λ is a hyperparameter used to balance the two loss items: $\mathcal{L}_{basic}$ and $\mathcal{L}_{avss}$. We conduct experiments to explore its impact on our semantic-aware optimization. As shown in Table 6, the model has the highest average performance when λ is set to 1. Therefore, this value is adopted as the optimal configuration.

Table 6: Impact of the hyperparameter λ. "*Avg.*" is the average result of all ten metrics. MM-Pyr [30] is used as the early audio-visual encoder.

λ	Segment-level					Event-level					*Avg.*
	A	V	AV	Type@AV	Event@AV	A	V	AV	Type@AV	Event@AV	
0.5	**64.8**	**67.8**	61.2	64.6	**63.7**	58.9	64.7	55.6	59.7	57.1	61.8
1.0	**64.8**	67.7	**61.8**	**64.8**	63.6	**59.2**	**64.9**	**56.5**	**60.2**	**57.4**	**62.1**
2.0	64.4	66.7	60.5	63.9	63.5	59.0	63.8	56.0	59.6	57.3	61.5

Table 7: Ablation study of the LEAP block. We determine which block's outputs are more suitable for final event prediction (denoted as "B-id"). "*Avg.*" is the average result of all ten metrics. MM-Pyr [30] is used as the early audio-visual encoder.

B-id	Segment-level					Event-level					*Avg.*
	A	V	AV	Type@AV	Event@AV	A	V	AV	Type@AV	Event@AV	
first	63.4	**67.1**	60.4	63.6	**62.8**	57.3	63.5	55.0	58.6	55.7	60.7
last	**63.7**	67.0	**61.3**	**64.0**	**62.8**	**58.2**	**63.9**	**56.2**	**59.5**	56.6	**61.3**
average	63.3	66.7	60.5	63.5	62.6	57.4	**63.9**	55.1	58.8	56.1	60.8

B Ablation study of LEAP block

In Table 1 of our main paper, we have established the optimal number (*i.e.*, 2) of LEAP blocks, we then explore which block's output is better suited for event predictions. We assess the outputs from the *first* block, the *last* block, and the *average* of these two blocks. As shown in 7, the best performance is obtained when using outputs from the last LEAP block. We speculate the cross-modal attention and enhanced label embedding are more discriminative at the last LEAP block.

We also conduct an ablation study which uses the learnable query of each event class to implement our LEAP method. Experimental results, as shown in Table 8, demonstrate that this strategy achieves competitive performance compared to using label embeddings extracted from the pretrained Glove model. The latter strategy (Glove) may provide more distinct semantics of different event classes, thereby facilitating model training in the initial phase and ultimately exhibiting slightly better performance.

Table 8: Ablation study on using learnable queries for label embedding in the proposed LEAP block.

Encoder	Setup	Segment-level					Event-level					*Avg.*
		A	V	AV	Type.	Eve.	A	V	AV	Type.	Eve.	
HAN	learnable	62.4	65.3	58.7	62.1	61.2	56.3	62.5	53.4	57.4	54.5	59.4
	glove	**62.7**	**65.6**	**59.3**	**62.5**	**61.8**	**56.4**	**63.1**	**54.1**	**57.8**	**55.0**	**59.8**
MM-Pyr	learnable	64.3	67.4	61.5	64.4	63.4	58.6	64.5	**56.7**	59.9	56.8	61.8
	glove	**64.8**	**67.7**	**61.8**	**64.8**	**63.6**	**59.2**	**64.9**	56.5	**60.2**	**57.4**	**62.1**

C Analysis of computational complexity

In Tables 2 and 3 of our main paper, we have demonstrated that our LEAP method can bring effective performance improvement particularly when combined with the advanced audio-visual encoder MM-Pyr [30]. Here, we further provide discussions on parameter overhead or computational complexity. 1) Our LEAP introduces more parameters than the typical decoding paradigm MMIL [23]. However, this increase is justified as MMIL merely utilizes several linear layers for event prediction, whereas our LEAP enhances the decoding stage with more sophisticated network designs and increases interpretability. By incorporating semantically distinct label embeddings of event classes, our LEAP involves increased cross-modal interactions between audio/visual and label text tokens. Consequently, our LEAP method inherently possesses more parameters than MMIL. 2) We further report the specific numbers of parameters and FLOPs of our LEAP-based model adopting the MM-Pyr as the audio-visual encoder. The total parameters of the entire model are 52.01M, while the parameters of our LEAP decoder are only 7.89M (15%). Similarly, the FLOPs of our LEAP blocks only account for 18.5% (146M v.s. 791M) of the entire model.

D More qualitative examples and analyses

We provide additional qualitative video parsing examples and analyses of our method. The MM-Pyr [30] is used as the early audio-visual encoder in this part.

The provided examples showcase the performance improvement and explainability of our proposed LEAP method compared to the typical decoding paradigm MMIL [23]. We discuss the details next.

As shown in Fig. 5, this video contains three overlapping events, *i.e.*, *cello*, *violin*, and *guitar*, occurring in both audio and visual modalities. Typical video parser MMIL [23] fails to correctly recognize the *cello* event for both audio and visual event parsing. In contrast, the proposed LEAP successfully identifies this event and provides more accurate predictions at the segment level. In the lower part of Fig. 5, we visualize the ground truth $\boldsymbol{Y}^m$, the cross-modal attention $\boldsymbol{A}^{lm}$ (intermediate output of our LEAP block, defined in Eq. 3 in our main paper), and the final predicted event probability $\boldsymbol{P}^m$, where $m \in \{a, v\}$ denotes the audio and visual modalities, respectively. It is noteworthy that the visualized $\boldsymbol{A}^{lm} \in \mathbb{R}^{C \times T}$ $(C = 25, T = 10)$ is processed by the softmax operation along the timeline as it goes through in LEAP block. $\boldsymbol{P}^m \in \mathbb{R}^{T \times C}$ is obtained through the raw cross-modal attention without the softmax operation and is activated by the sigmoid function. We show the transpose of $\boldsymbol{P}^m$ in the figure. In this video example, all three events generally appear in all the video segments. Therefore, their corresponding label embeddings exhibit similar cross-modal (audio/visual-label) attention weights for all the temporal segments, as highlighted by the red rectangular frames in Fig. 5. In this way, the label embeddings of these three events can be enhanced by aggregating relevant semantics from all the highly matched temporal segments and then are used to predict correct event classes. Moreover, the visualization of $\boldsymbol{P}^m$ indicates that our LEAP effectively learns meaningful cross-modal relations between each segment and each label embedding of audio/visual events, yielding predictions similar to the ground truth $\boldsymbol{Y}^m$.

A similar phenomenon can also be observed in Fig. 6. Both typical video decoder MMIL and our LEAP correctly localize the visual event *dog*. However, MMIL incorrectly recognizes most of the video segments as containing the audio events *speech* and *dog*. In contrast, the proposed LEAP provides more accurate segment-level predictions for audio event parsing. As verified by the visualization of the cross-modal attention $\boldsymbol{A}^{lm}$, the label embeddings of *speech* and *dog* classes mainly have large similarity weights for those segments that genuinely contain the corresponding events (marked by the red box). This distinction allows our LEAP-based method to better differentiate the semantics of various events and provide improved segment-level predictions.

In summary, these visualization results provide further evidence of the advantages of our LEAP method in addressing overlapping events, enhancing different event recognition, and providing explainable results.

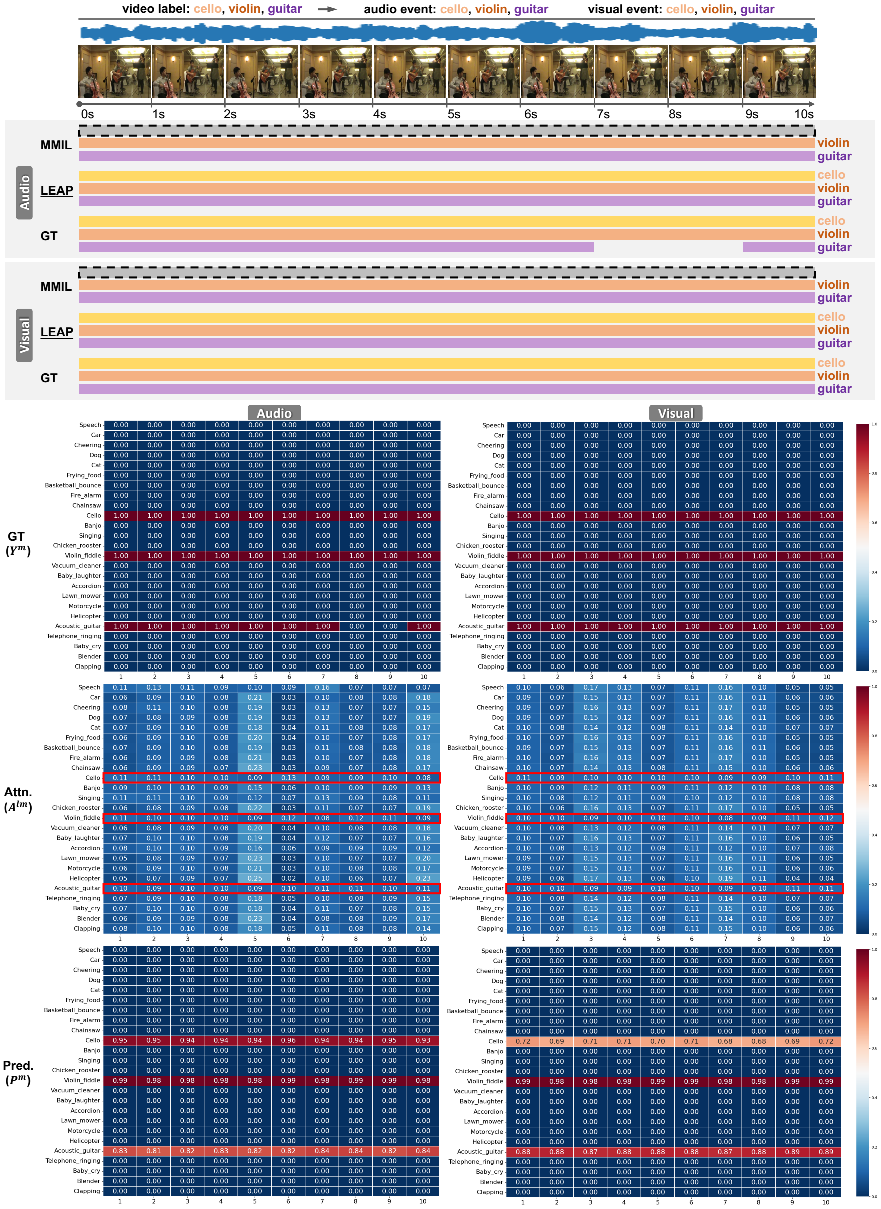

Fig. 5: More qualitative video examples of audio-visual video parsing. Best view in color and zoom in.

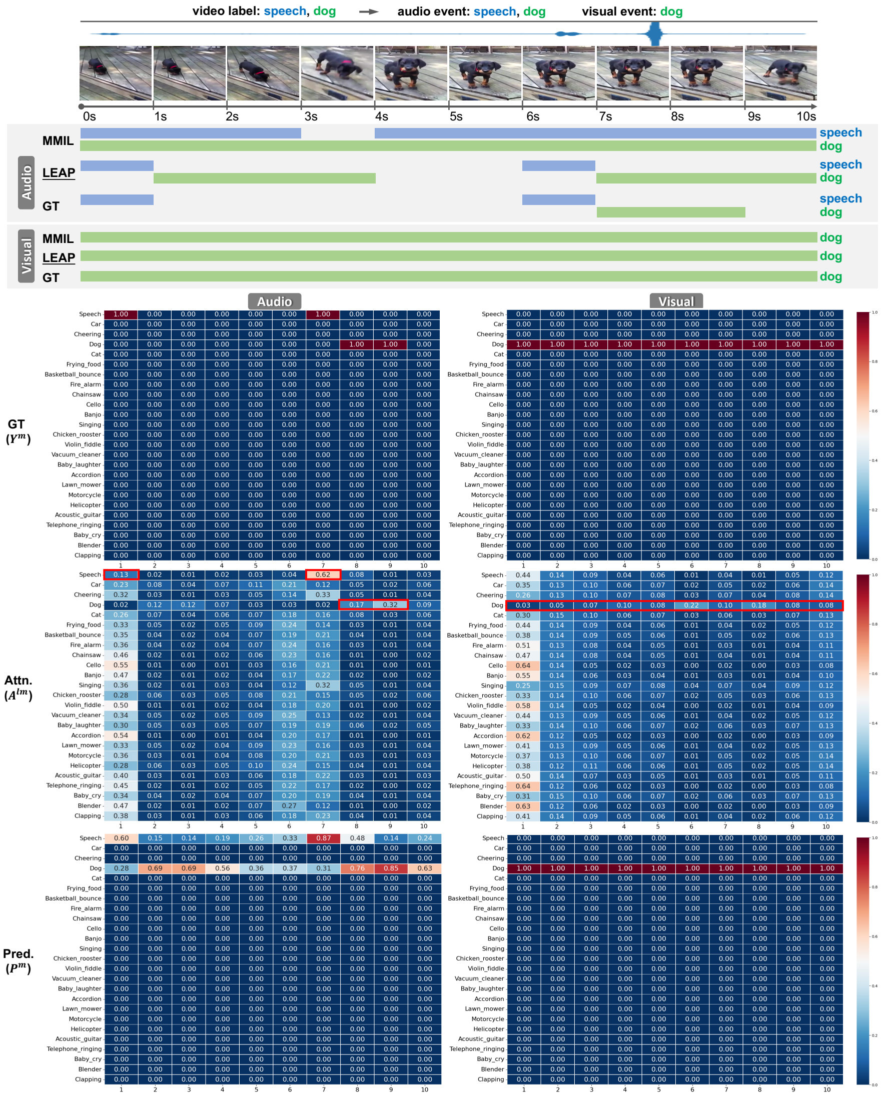

Fig. 6: More qualitative video examples of audio-visual video parsing. Best view in color and zoom in.