JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception

Junyu Gao, Mengyuan Chen, Changsheng Xu

State Key Laboratory of Multimodal Artificial Intelligence Systems
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Peng Cheng Laboratory

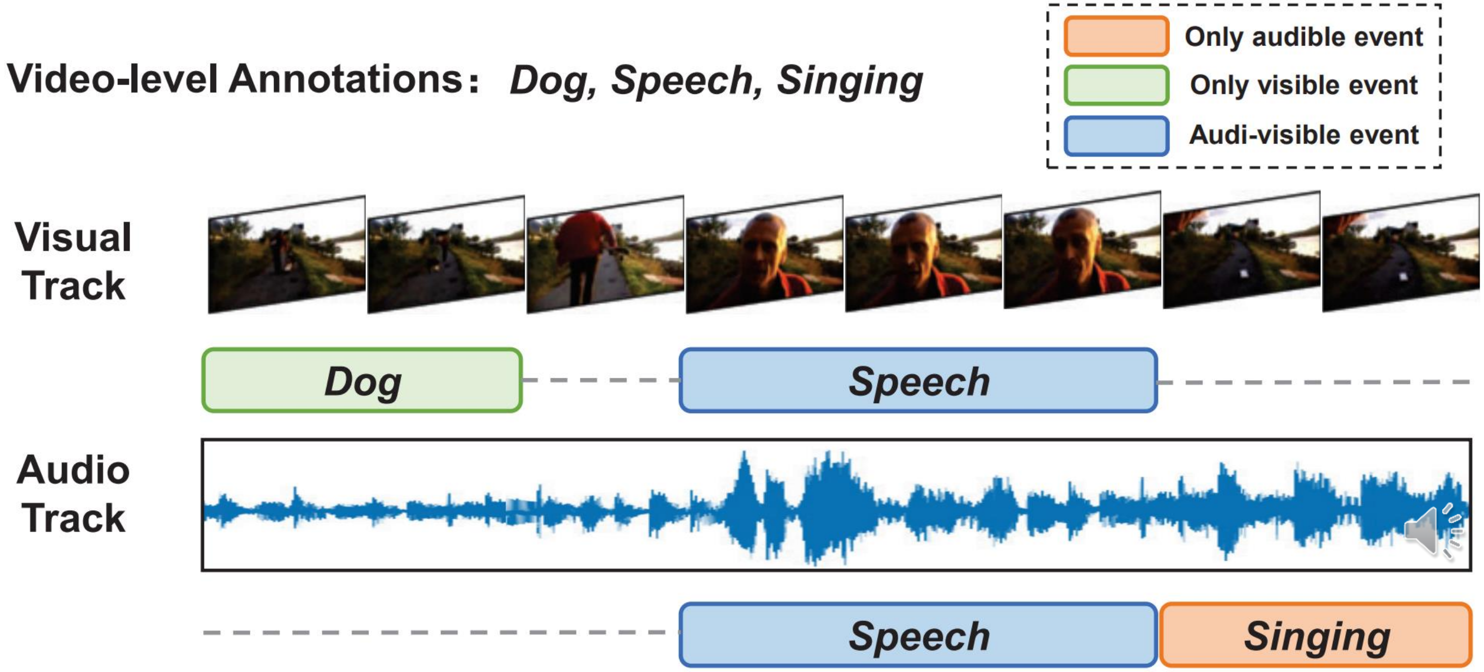
2023-06

THU-AM-224

Introduction

■ Weakly-supervised Audio-Visual Event Perception

- With **only video-level annotations**, weakly-supervised audio-visual event perception (WS-AVEP) aims to predict the temporal boundaries of various **only audible** (in orange), **only visible** (in green), or **audi-visible** (in blue) events in a video.



Introduction

■ Two main pipelines of WS-AVEP

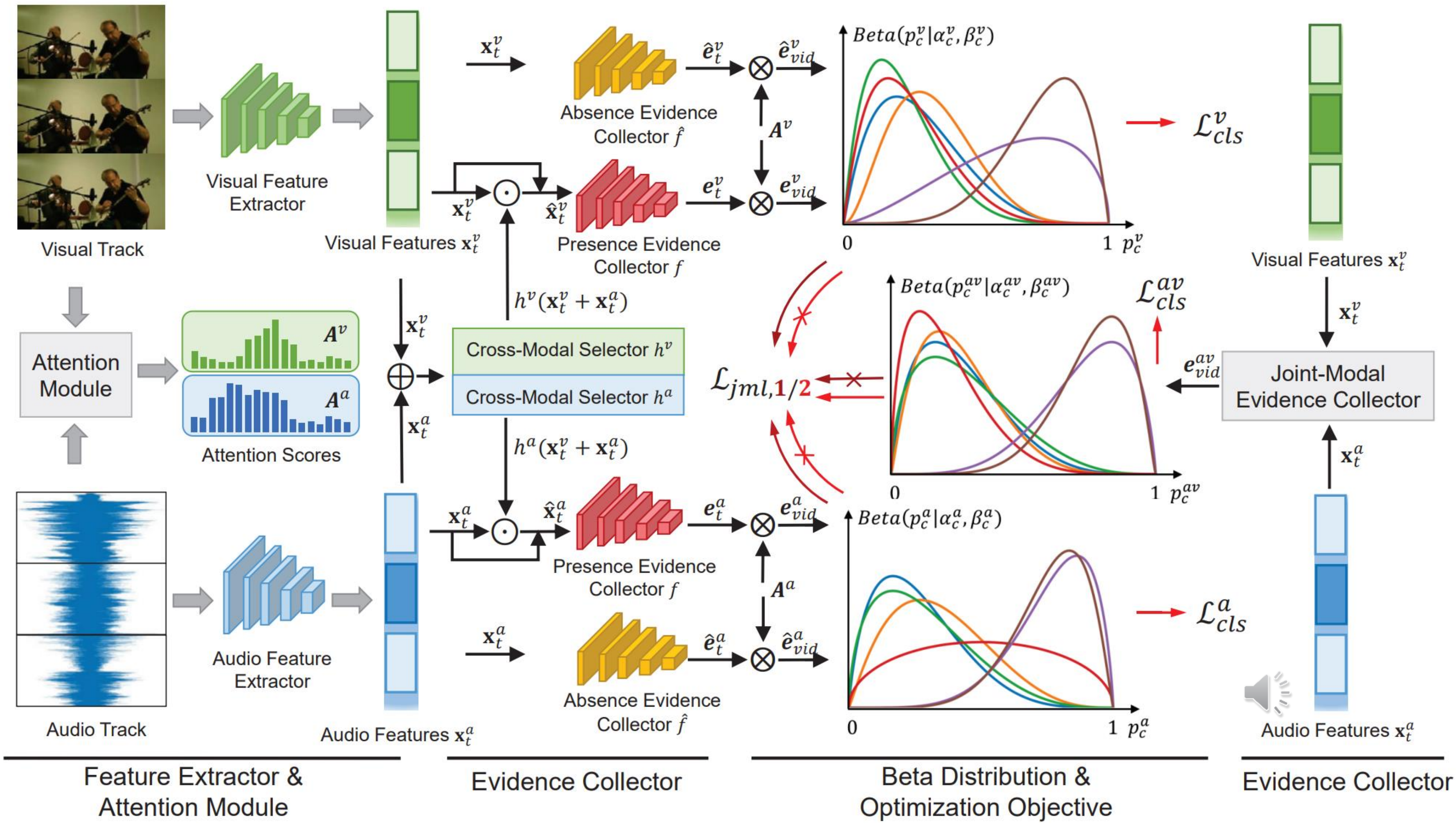
- **AVE**: Events are all **simultaneously audible and visible**.
- **AVVP**: Events are categorized into **only audible, only visible, or audi-visible** ones.
- State-of-the-arts methods can only achieve significant performance in one single WS-AVEP setting, showing that current methods are **in a dilemma of making full use of both uni-modal and cross-modal information.**

Method / Task	CMBS [61]	JoMoLD [6]	**Ours**
AVVP [52]	51.7	**57.3**	**60.1**
AVE [53]	**74.2**	71.8	**74.8**

The modality itself should provide ample presence evidence of this event, while the other complementary modality is encouraged to afford the absence evidence as a reference signal.

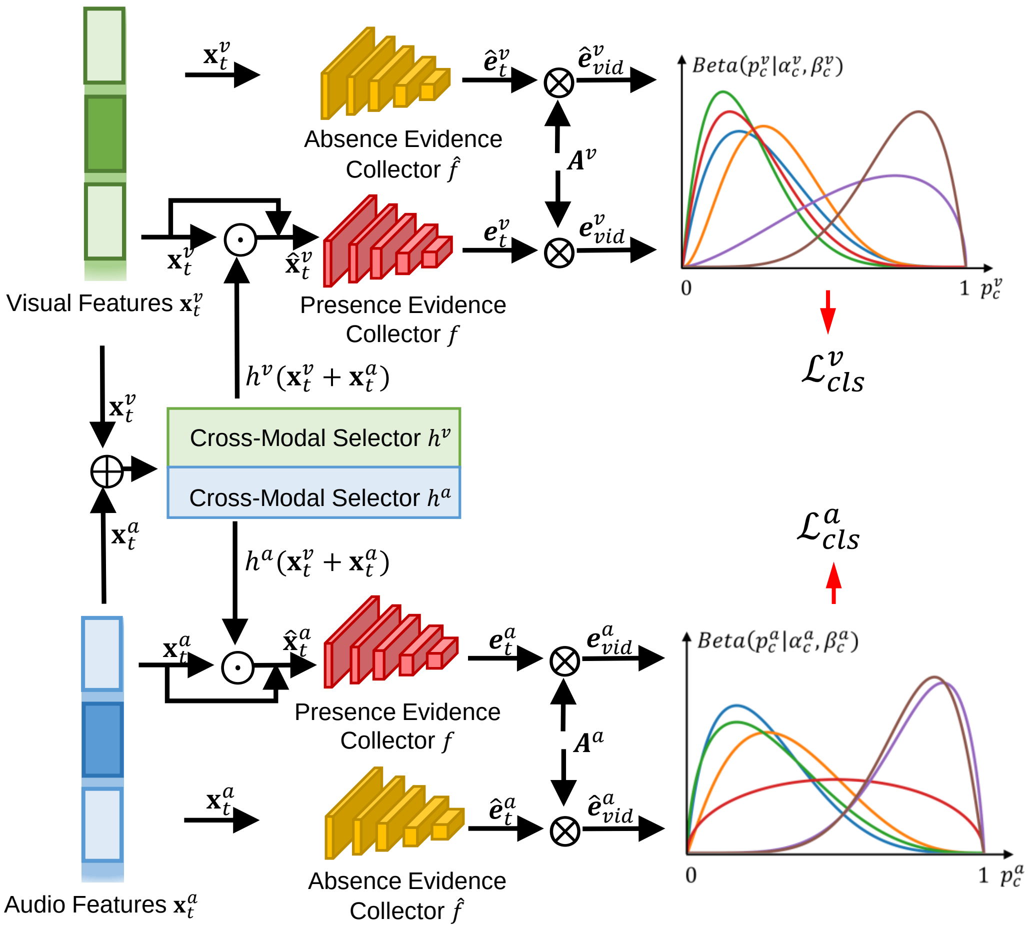
Method

■ Cross-modal Presence-absence Evidence Learning



Method

■ Presence-absence Evidence Collector



➢ **Presence evidence**

$$e_{t,c}^m = g\left(f_c(\mathbf{x}_t^m; \boldsymbol{\theta}_1)\right)$$

➢ **Absence evidence**

$$\hat{e}_{t,c}^m = g\left(\hat{f}_c(\hat{\mathbf{x}}_t^m; \boldsymbol{\theta}_2)\right),$$

$$\hat{\mathbf{x}}_t^m = \mathbf{x}_t^m \odot \left(h^m(\mathbf{x}_t^m + \mathbf{x}_t^{\hat{m}}; \boldsymbol{\theta}_3^m) + \mathbf{1}\right)$$

➢ **Loss function**

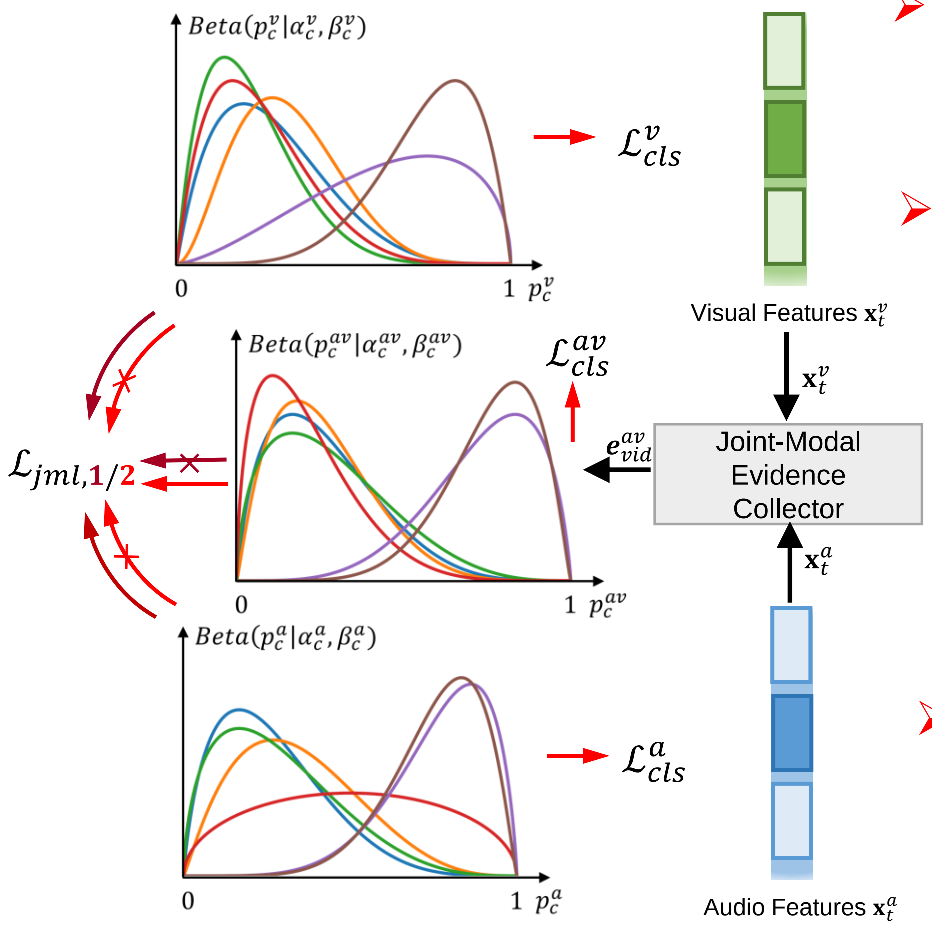
$$\mathrm{Beta}(p_c|\alpha_c, \beta_c) = \frac{1}{B(\alpha_c, \beta_c)} p_c^{\alpha_c - 1}(1 - p_c)^{\beta_c - 1}$$

$$\mathcal{L}_{cls}^m = \int \left[\sum_{c=1}^{C} -y_c^m \log(p_c)\right] \mathrm{Beta}(p_c|\alpha_c, \beta_c) dp$$

$$= \sum_{c=1}^{C} \left[\psi(\alpha_c + \beta_c) - \psi\left(y_c^m \alpha_c + (1 - y_c^m)\beta_c\right)\right]$$

Method

■ Joint-modal Mutual Learning



- **Video-level presence/absence evidence**

$$e_{vid,c}^{av}, \hat{e}_{vid,c}^{av} = \sum_t A_{t,c}^{av} \cdot g(f_c^{av}(\mathbf{x}_t^a + \mathbf{x}_t^v; \boldsymbol{\theta}_4))$$

- **Cross-modal fusion**

$$p_c^m = \frac{e_c^m + 1}{e_c^m + \hat{e}_c^m + 2}, \quad u_c^m = \frac{2}{e_c^m + \hat{e}_c^m + 2}$$

$$\{u_c^{uni}, p_c^{uni}\} = \delta(c)\{u_c^a, p_c^a\} + (1 - \delta(c))\{u_c^v, p_c^v\}$$

$$\delta(c) = \begin{cases} 1, & p_c^a > p_c^v, y_c = 1, \\ 0, & p_c^a \le p_c^v, y_c = 1, \\ 1/2, & y_c = 0. \end{cases}$$

- **Loss function for mutial learning**

$$\mathcal{L}_{jml,1} = \sum_c (1 - u^{av})\left(1 - u_c^{uni}\right) * l\left(s(p_c^{av}), p_c^{uni}\right)$$

$$\mathcal{L}_{jml,2} = \sum_c u^{av}\left(1 - u_c^{uni}\right) * l\left(p_c^{av}, s(p_c^{uni})\right),$$

Experiments

■ Evaluation on AVVP / AVE / AVEP

Table 2. AVVP performance comparison with existing methods on the LLP dataset.

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
AVE [53], ECCV2018	49.9	37.3	37.0	41.4	43.6	43.6	32.4	32.6	36.2	37.4
AVSDN [29], ICASSP2019	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [52], ECCV2020	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
CVCMS [30], NeurIPS2021	60.8	63.5	57.0	60.5	59.5	53.8	58.9	49.5	54.0	52.1
MA [58], CVPR2021	59.8	57.5	52.6	56.6	56.6	52.1	54.4	45.8	50.8	49.4
DHHN [22], MM2022	61.4	63.4	56.8	60.5	59.5	54.6	60.8	51.1	55.5	53.3
MM-Pyramid [65], MM2022	61.1	60.3	55.8	59.7	59.1	53.8	56.7	49.4	54.1	51.2
CMBS* [61], CVPR2022	60.2	54.3	50.0	54.8	55.7	51.1	50.8	43.7	48.5	48.3
JoMoLD [6], ECCV2022	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5
CMPAE(Ours)	**64.2 (+2.9)**	**66.4 (+2.6)**	**59.2 (+2.0)**	**63.3 (+2.5)**	**62.8 (+2.9)**	**56.6 (+2.7)**	**63.7 (+3.8)**	**51.8 (+2.2)**	**57.4 (+2.9)**	**55.7 (+3.2)**

Table 3. AVE performance comparison.

Methods	Accuracy(%)
AVEL [53], ECCV2018	66.7
AVRB [47], WACV2020	68.9
CMRAN [62], MM2020	72.9
PSP [70], CVPR2021	73.5
CMAN [63], AAAI2022	70.4
MM-Pyramid [65], MM2022	73.2
CMBS [61], CVPR2022	74.2
DPNet [48], ECCV2022	74.5
CMBS [61], fully-supervised	79.3
JoMoLD* [6], ECCV2022	71.8
CMPAE(Ours)	**74.8**

* denotes the reproduced results.

Table 4. AVEP performance comparison with existing methods.

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
CMBS [61], CVPR2022	58.0	56.2	52.3	55.5	54.8	51.5	53.6	46.4	50.5	49.4
JoMoLD [6], ECCV2022	60.6	58.9	54.5	58.0	57.7	53.6	55.8	48.6	52.7	51.0
CMPAE(Ours)	**64.1 (+3.5)**	**64.4 (+5.5)**	**58.8 (+4.3)**	**62.4 (+4.4)**	**62.2 (+4.5)**	**57.2 (+3.6)**	**61.9 (+6.1)**	**52.3 (+3.7)**	**57.1 (+4.4)**	**55.6 (+4.6)**

Experiments

- **Ablation study**

Table 5. Ablation studies of our method.

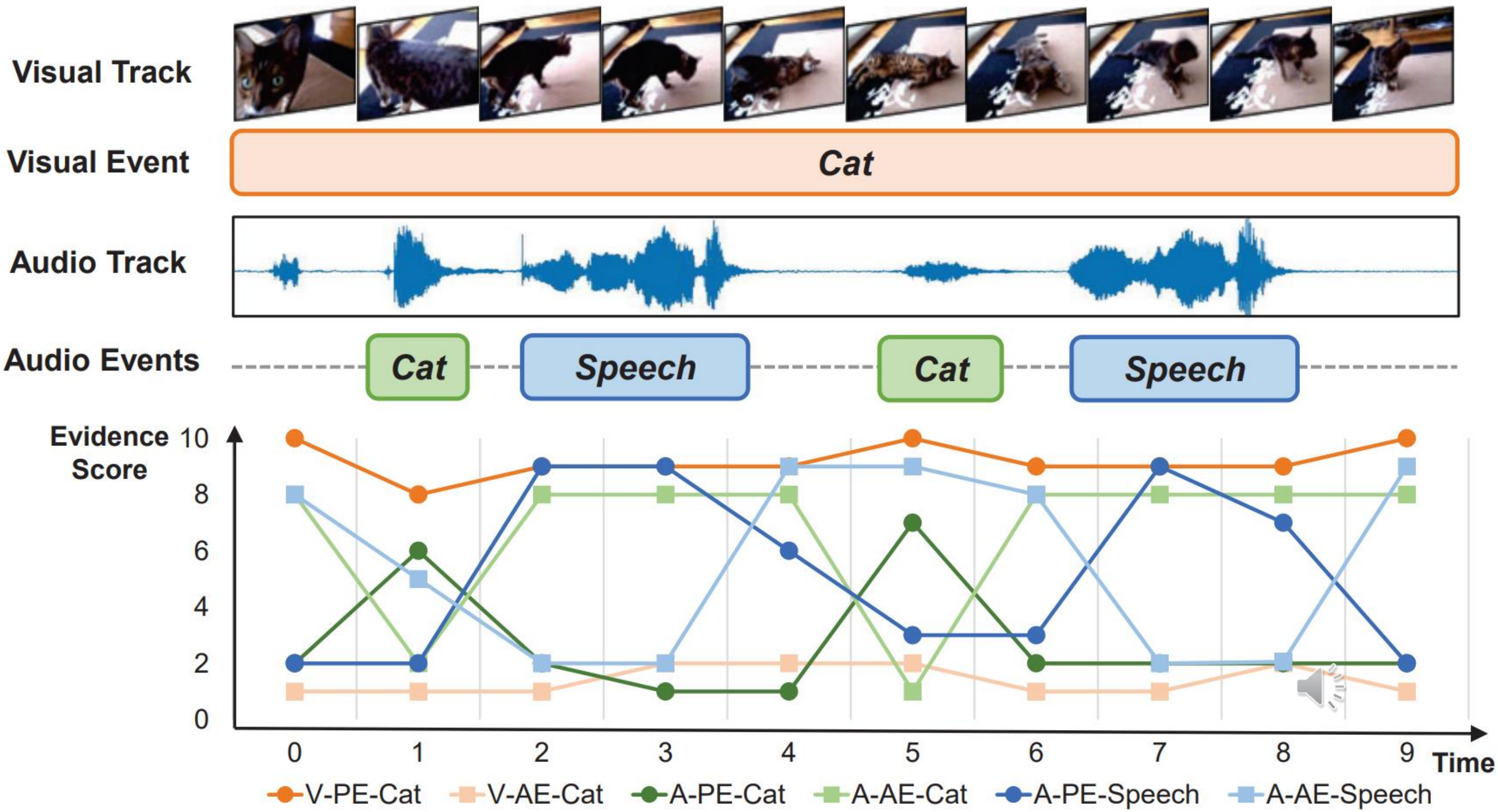
EDL	PAEC	JML	Seg-level Type AVVP	Seg-level Type AVEP	Eve-level Type AVVP	Eve-level Type AVEP
✗	✗	✗	60.8	58.0	54.5	52.7
✓	✗	✗	61.0	58.9	54.9	53.8
✓	✓	✗	61.9	61.5	56.1	55.9
✓	✗	✓	61.4	60.8	55.3	54.6
✓	✓	✓	**63.3**	**62.4**	**57.4**	**57.1**

Table 6. In-depth analysis of our proposed PAEC and JML.

Models	Seg-level Type AVVP	Seg-level Type AVEP	Eve-level Type AVVP	Eve-level Type AVEP
both uni-modal	61.2	60.9	55.2	54.4
both cross-modal	61.7	61.3	55.7	55.9
exchange uni/cross	62.1	61.6	56.4	56.0
w/o u^{av}	62.2	61.8	56.5	56.3
w/o u^{uni}	62.0	61.7	56.4	56.0
w/o $\delta(c)$	62.5	61.8	56.3	56.5
CMPAE	**63.3**	**62.4**	**57.4**	**57.1**

Experiments

- **Visualization Analysis**



Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception

Code & Model

Any problem, please feel free contact the primary author:

Junyu Gao
junyu.gao@nlpr.ia.ac.cn