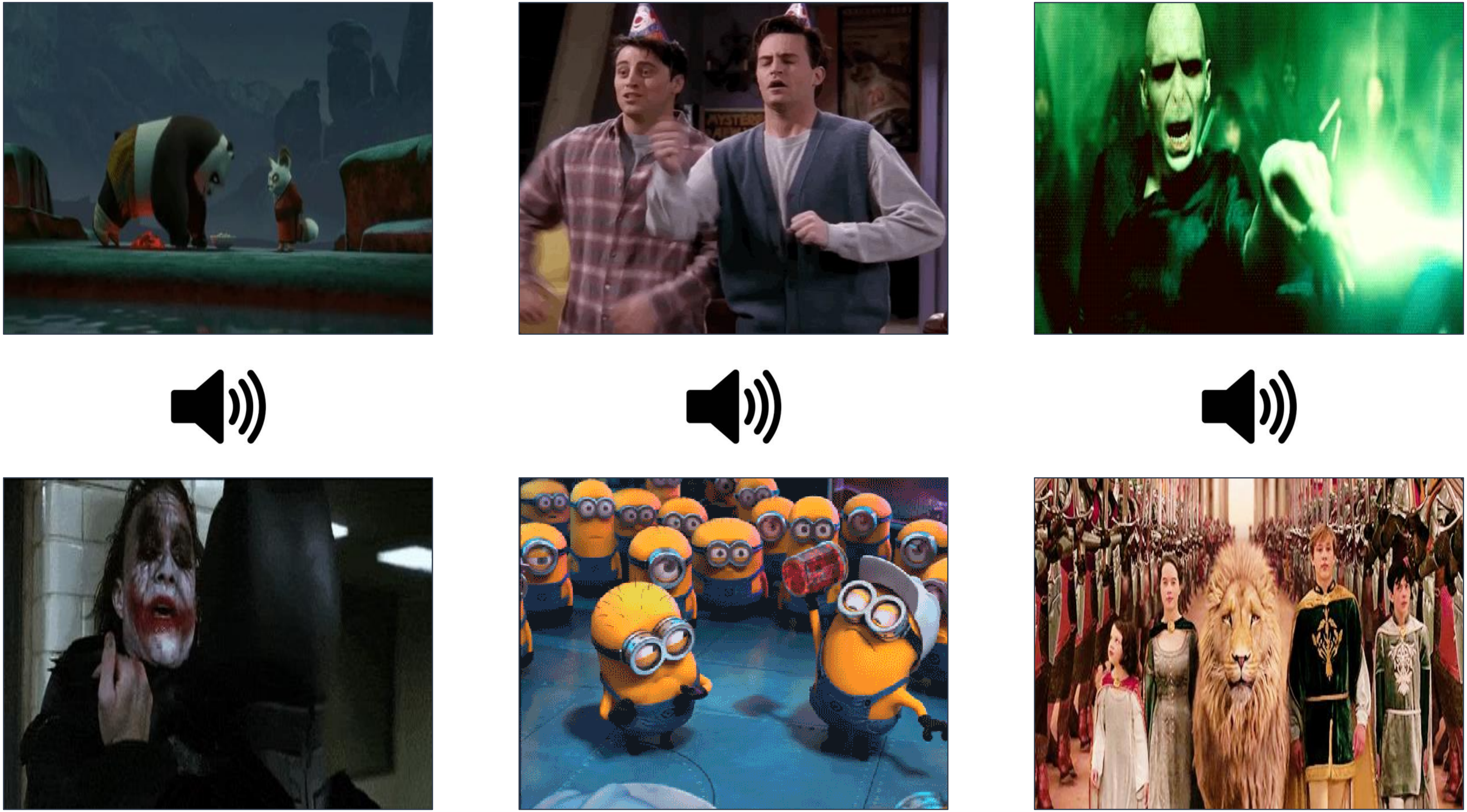
WHAT MAKES THE SOUND?: A DUAL-MODALITY INTERACTING NETWORK FOR AUDIO-VISUAL EVENT LOCALIZATION

Janani Ramaswamy

Dept. of CS & E, IIT Madras, Chennai-600036, India



Can’t machines mimic humans in using both audio and video for decision making?

CHALLENGES

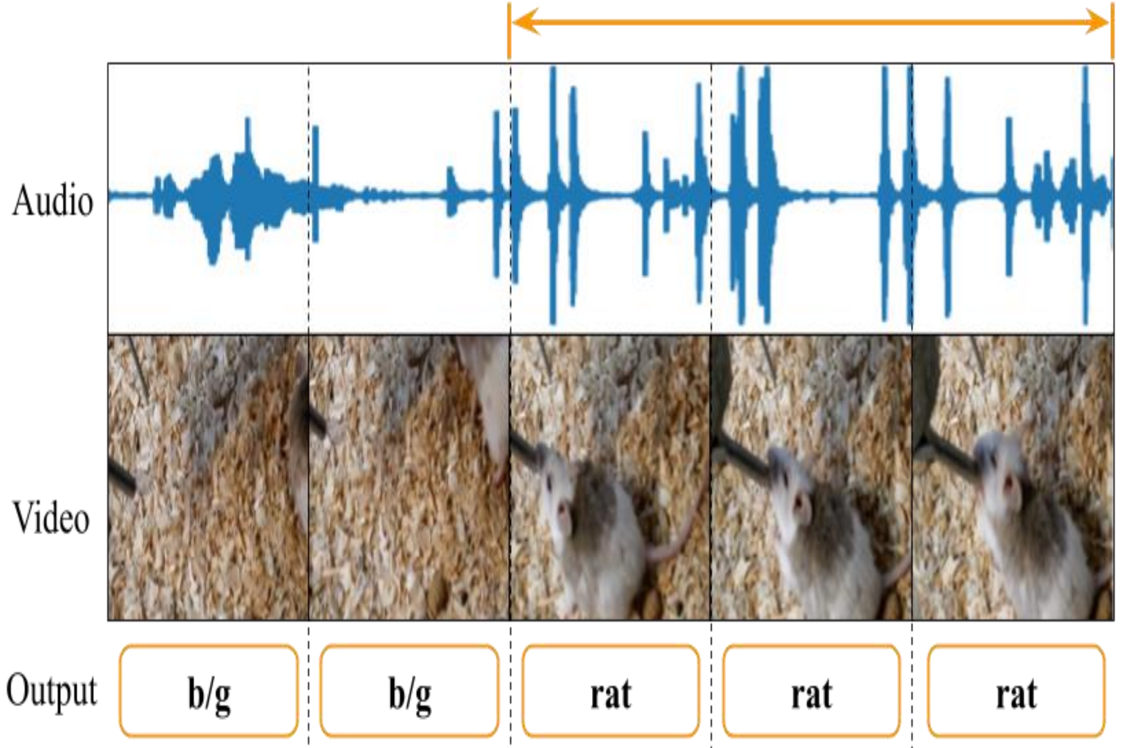
- Audio may not always be in perfect sync with the video
- Presence of ambient sound like breeze
- Object making the sound being momentarily occluded in the video
- Obtaining the semantics is less direct in case of audio[1].

1. R. Arandjelovic and A. Zisserman, Objects that sound, ICCV 2017.

AUDIO-VISUAL EVENT LOCALIZATION

Input Video

Supervised event localization:
Training: event label given for every segment
Testing: predict event category for every segment



Weakly-Supervised event localization:
Training: event label given for whole video
Testing: predict event category for every segment

APPLICATIONS

- Audio-based video captioning
- Audio-based video segmentation
- Surveillance
- Enhanced scene understanding

EXISTING WORKS

Tian et al. ECCV 2018 [1]

- Audio-visual event localization in unconstrained videos
- Audio-Visual Event (AVE) dataset

Wu et al. ICCV 2019 [3]

- Dual Attention Matching (DAM)
- Encodes temporal co-occurrence between auditory and visual signals

Lin et al. ICASSP 2019 [2]

- Audio-Visual seq2seq dual n/w (AVSDN)
- learns global and local event info in seq2seq manner

Ramaswamy & Das WACV 2020 [4]

- Spatial & Segment-wise attention using two novel blocks
- A novel loss function for unsupervised sound localization

1. Y. Tian, J. Shi, B. Li, Z. Duan and C. Xu, Audio-visual event localization in unconstrained videos, ECCV 2018.
2. Y.-B. Lin, Y.-J. Li and Y.-C. F. Wang, Dual-modality seq2seq network for audio-visual event localization, ICASSP 2019.
3. Y. Wu, L. Zhu, Y. Yan and Y. Yang, Dual Attention Matching for Audio-Visual Event Localization, ICCV 2019.
4. J. Ramaswamy and S. Das, See the Sound, Hear the Pixels, WACV 2020.

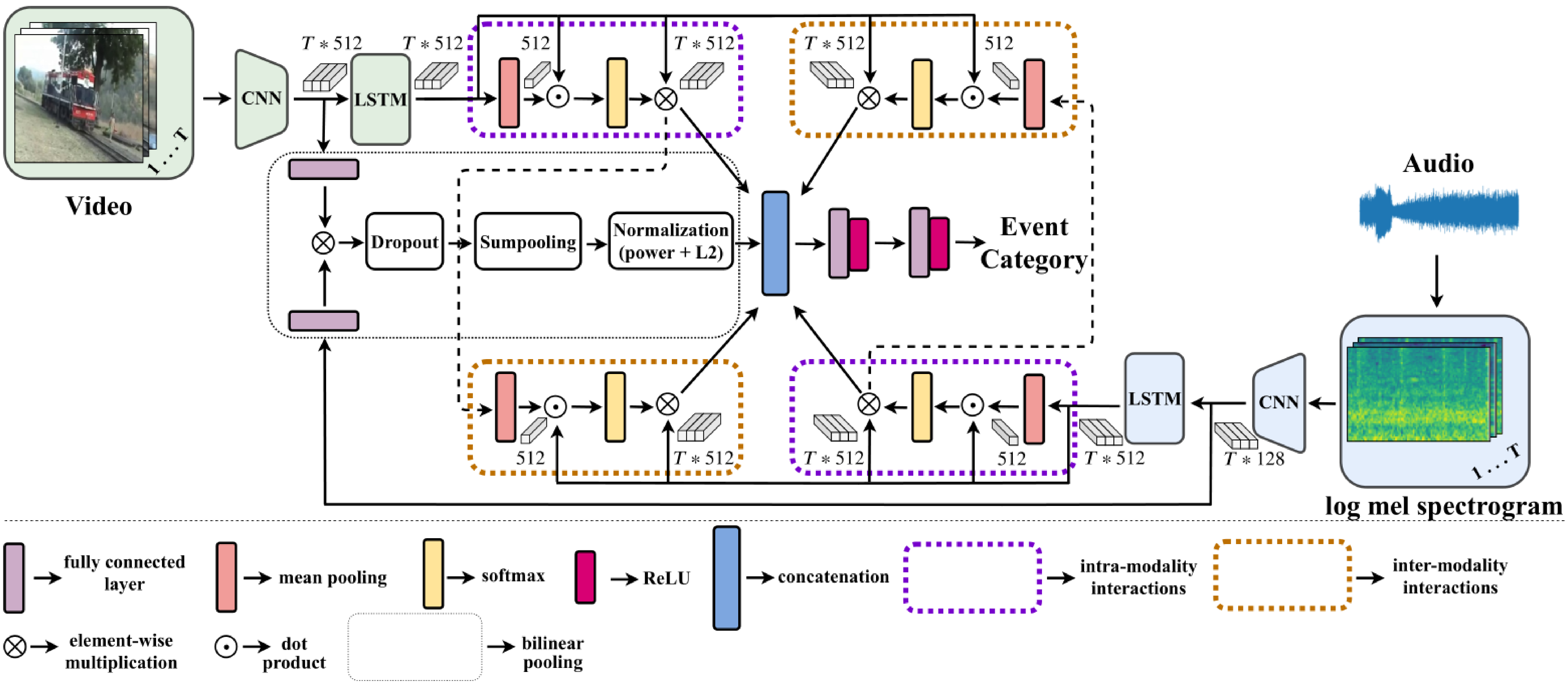
MAJOR CONTRIBUTIONS

Audio-Visual Interacting Network (AVIN) for fully & weakly supervised audio-visual event localization

A novel audio-visual fusion that captures the inter and intra modality interactions using local and global information from the two modalities

Our method significantly outperforms the existing state-of-the-art methods

PROPOSED ARCHITECTURE



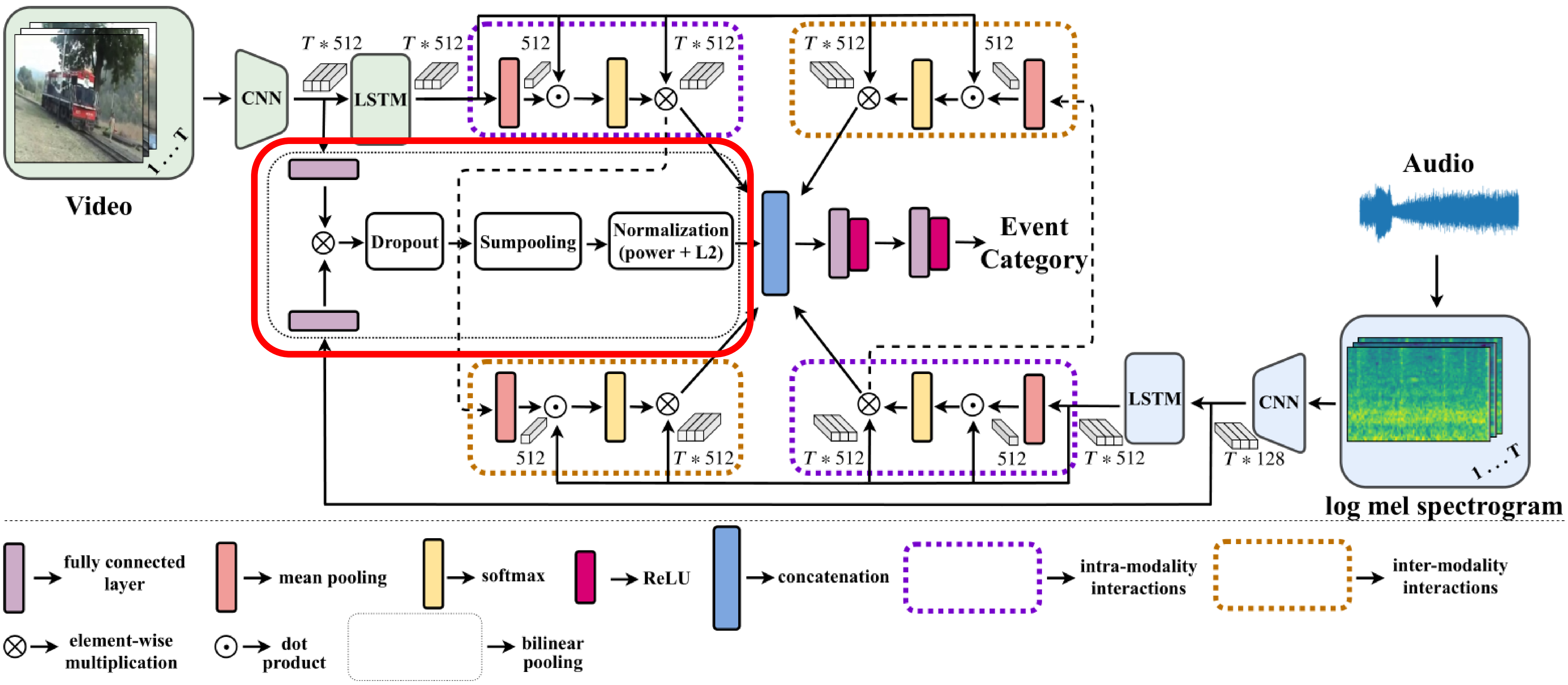
Audio-Visual Interacting Network (AVIN)

- **Feature Extraction**: Let $F_t^a \in \mathbb{R}^{d_a}$ and $F_t^v \in \mathbb{R}^{d_v}$ denote the audio and visual features extracted using CNNs. Here, d_a and d_v refer to the dimension of audio and visual features respectively.

- **Modeling temporal dependency**: The features $\{F_t^a, F_t^v\}_{t=1}^{\mathcal{T}}$ extracted from the CNNs are then fed to two different LSTMs, the result of which is denoted as $\{h_t^a, h_t^v\}_{t=1}^{\mathcal{T}}$.

- $\mathcal{T}$ here denotes the number of non-overlapping segments ($= 10$ in our case) that each video is split into.

CAPTURING BILINEAR INTERACTIONS



Bilinear Pooling for audio-visual fusion

- Consider a multi-modal bilinear model :

$$\tilde{z}_t = {F_t^v}^T W_i F_t^a \quad \text{-------------- (1)}$$

where, $W_i \in \mathbb{R}^{d_v \times d_a}$ is the projection matrix and $\tilde{z}_t$ is a scalar.

- To get a p-dimensional output, we use $W = [W_1, \dots, W_p] \in \mathbb{R}^{d_v \times d_a \times p}$
- But this leads to a large number of parameters and high computational cost.
- Multi-modal Factorized Bilinear (MFB) Pooling[1] factorizes W into two low-rank matrices:

$$\tilde{z}_t = Sumpooling(U^T F_t^v \circ V^T F_t^a, q) \quad \text{-------------- (2)}$$

- Applying power and L2 normalization:

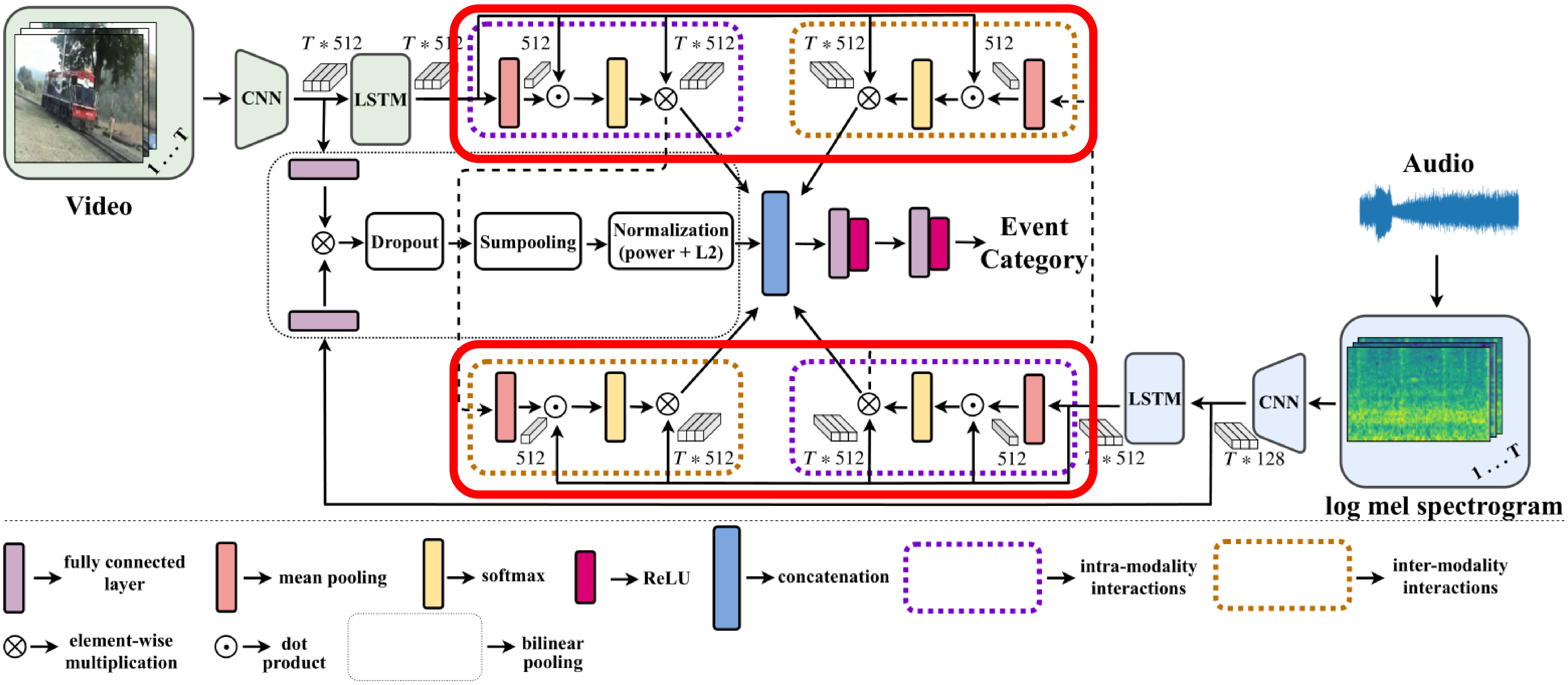
$$z_t' = sign(\tilde{z}_t)|\tilde{z}_t|^{0.5}; z_t = z_t'/||z_t'|| \quad \text{-------------- (3)}$$

Where, $U \in \mathbb{R}^{d_v \times (qp)}$ and $V \in \mathbb{R}^{d_a \times (qp)}$ are the two low rank matrices.

$\circ$ refers to the Hadamard product and q represents the latent dimensionality.

1. Yu et al., Multi-modal factorized bilinear pooling with co-attention learning for visual question-answering, ICCV 2017.

Capturing inter and intra modality interactions



Capturing inter and intra modality interactions

- To get a better idea about the amount of synchronization present between the two modalities, the global information also needs to be considered.
- We use self and collaborative attention[1] to capture intra and inter modality interactions.
- **Intra-modality interactions:**

$$s_t^a = Softmax\left(h_t^a \odot \bar{h}_{ave}^a\right) \otimes h_t^a \qquad \text{------- (4)}$$

$$s_t^v = Softmax\left(h_t^v \odot \bar{h}_{ave}^v\right) \otimes h_t^v \qquad \text{------- (5)}$$

- **Inter-modality interactions:**

$$c_t^a = Softmax(h_t^a \odot \bar{s}_{ave}^v) \otimes h_t^a \qquad \text{------- (6)}$$

$$c_t^v = Softmax(h_t^v \odot \bar{s}_{ave}^a) \otimes h_t^v \qquad \text{------- (7)}$$

where,

h_t^a, h_t^v - temporally encoded features from LSTMs

$\bar{h}_{ave}^a, \bar{h}_{ave}^v$ - outputs of mean pooling applied on h_t^a, h_t^v

s_t^a, s_t^v - features encoded with **intra**-modality interactions

c_t^a, c_t^v - features encoded with **inter**-modality interactions

$\bar{s}_{ave}^a, \bar{s}_{ave}^v$ - outputs of mean pooling applied on s_t^a, s_t^v

$\odot$ - dot product

$\otimes$ - element-wise multiplication

1. Zhang et al., Scan: Self-and-collaborative attention network for video person re-identification, TIP 2019.

DATASET USED

Audio-Visual Event (AVE) Dataset[1]

- 4143 videos (min 2s long event; max 10s long event)
- 28 event categories
- Minimum of 60 and maximum of 188 videos in each category
- Labels available video-wise as well as segment-wise (i.e., temporally labeled) with audio-visual event boundaries.

1. Y. Tian, J. Shi, B. Li, Z. Duan and C. Xu, Audio-visual event localization in unconstrained videos, ECCV 2018.

RESULTS (PERFORMANCE COMPARISON IN %)

Method	Sup. Acc.	W-Sup. Acc.
Audio	62.3	57.0
Visual	57.4	53.8
AVE[1]	72.7	66.7
AVSDN[2]	72.8	66.5
DAM[3]	74.5	-
Ramaswamy & Das[4]	74.8	68.9
AVIN (Ours: Aud + Vis)	**75.2**	**69.4**

1. Y. Tian, J. Shi, B. Li, Z. Duan and C. Xu, Audio-visual event localization in unconstrained videos, ECCV 2018.
2. Y.-B. Lin, Y.-J. Li and Y.-C. F. Wang, Dual-modality seq2seq network for audio-visual event localization, ICASSP 2019.
3. Y. Wu, L. Zhu, Y. Yan and Y. Yang, Dual Attention Matching for Audio-Visual Event Localization, ICCV 2019.
4. J. Ramaswamy and S. Das, See the Sound, Hear the Pixels, WACV 2020.

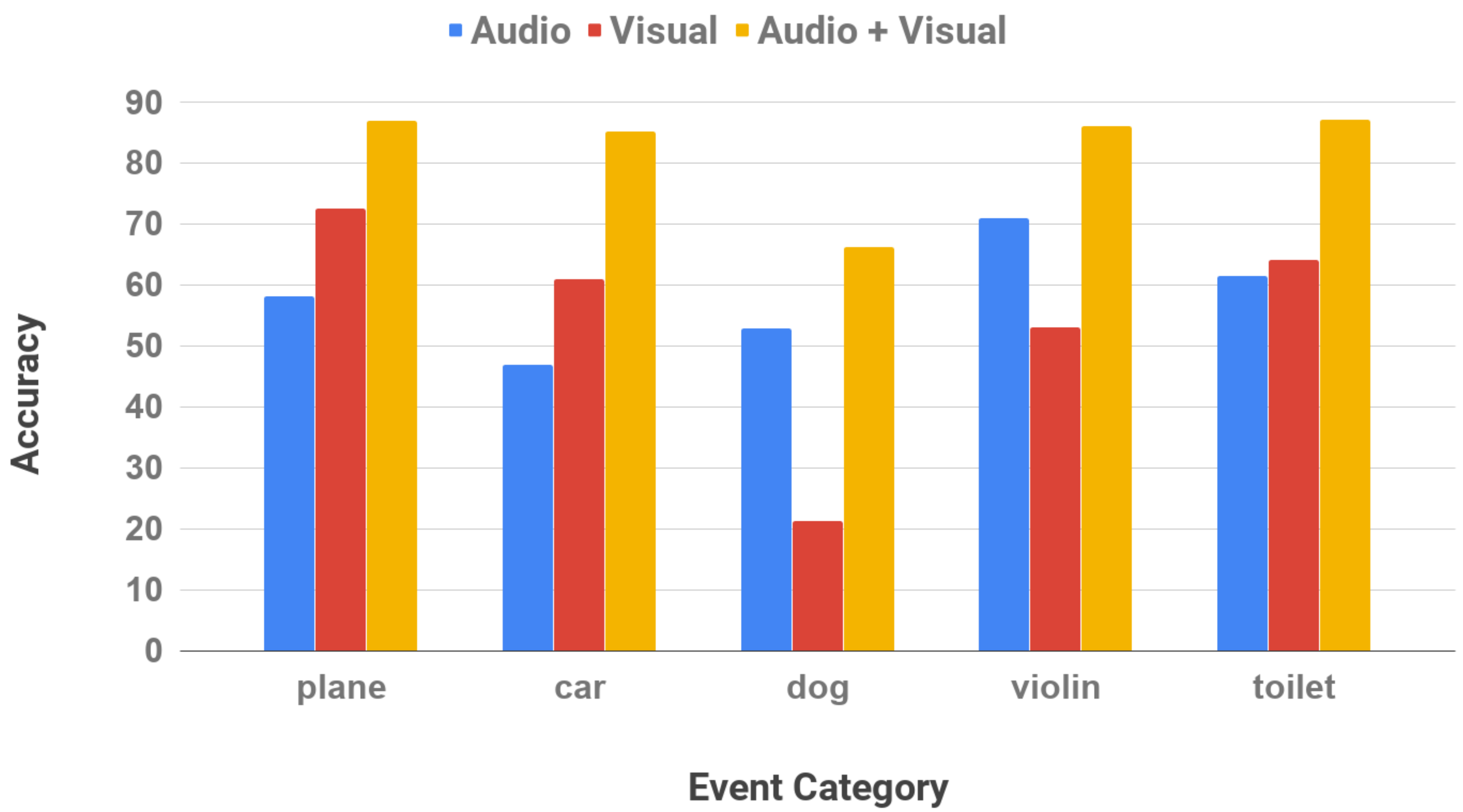
RESULTS (DIFFERENT FUSION STRATEGIES)

Fusion Strategy	Sup. Acc.	W-Sup. Acc.
Element-wise multiplication	60.3	55.1
Element-wise addition	63.4	58.2
Concatenation + FC	65.7	60.3
AVIN (Ours)	**75.2**	**69.4**

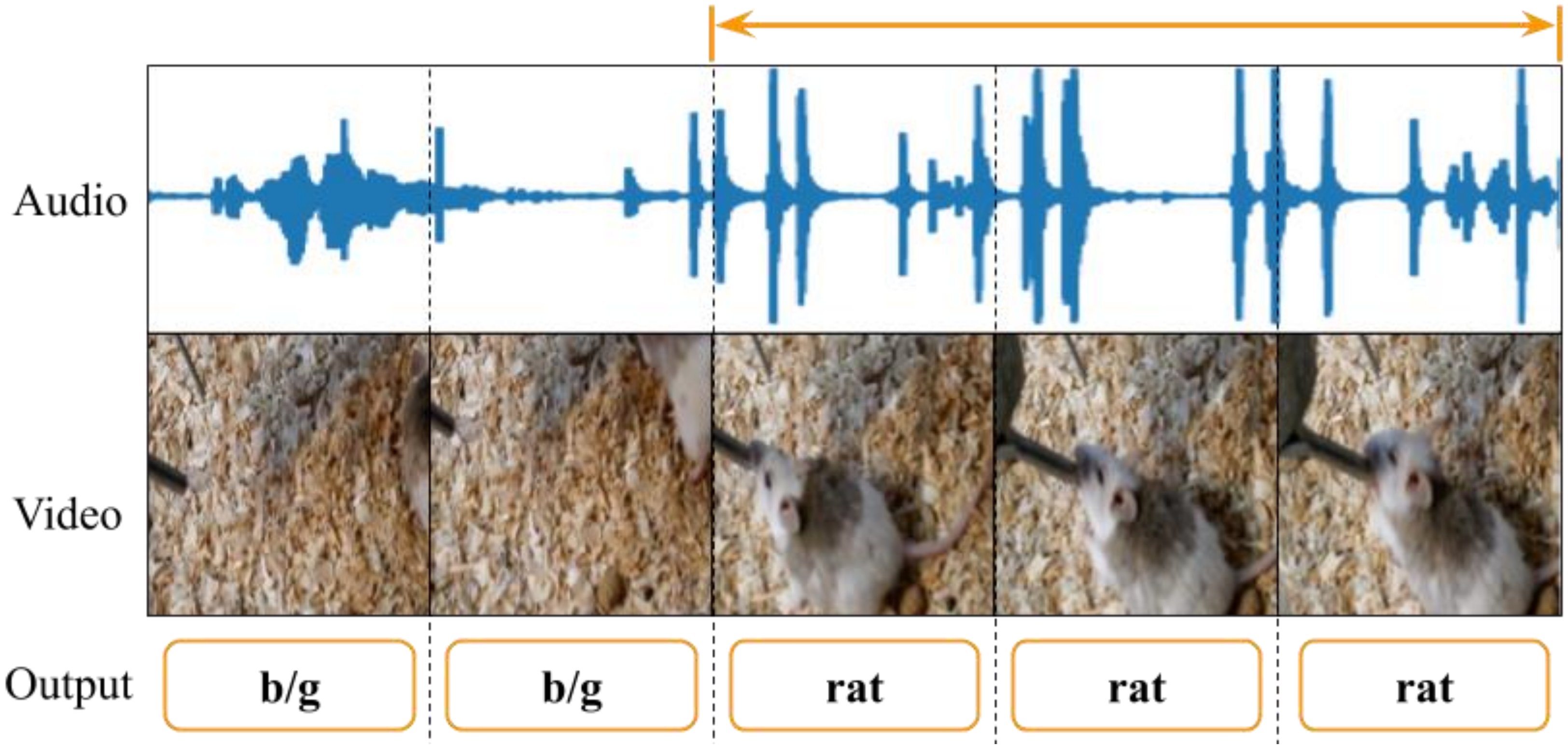
ABLATION STUDY

Model	Sup. Acc.	W-Sup. Acc.
Only LSTM	70.1	63.8
Only MFB[1]	71.4	66.7
LSTM + intra-mod	71.2	65.4
LSTM + intra + inter-mod	73.5	67.9
LSTM + MFB + intra+ inter-mod	**75.2**	**69.4**

1. Z. Yu, J. Yu, J. Fan and D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question-answering, ICCV 2017.



Bar chart depicting accuracies of a few selected event categories for supervised event localization task



Output of a few segments shown for our proposed method of supervised event localization, given an input video.

THANK YOU!

JANANI RAMASWAMY

(Research Scholar, IIT Madras)

Visualization and Perception Lab – www.cse.iitm.ac.in/~vplab