Modality-Independent Teachers Meet Weakly-Supervised Audio-Visual Event Parser

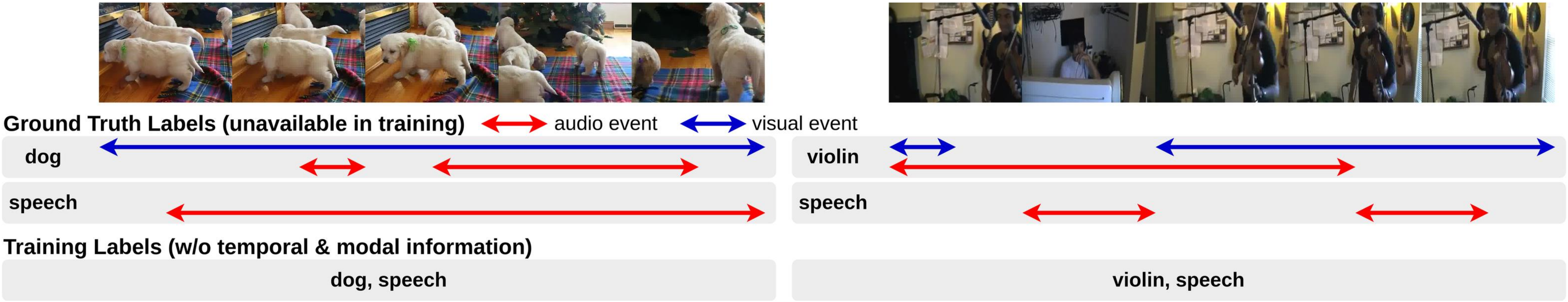
Yung-Hsuan Lai[1], Yen-Chun Chen[2], Yu-Chiang Frank Wang[1,3]

[1]National Taiwan University [2]Microsoft [3]NVIDIA

NeurIPS 2023

Audio-Visual Video Parsing (AVVP)

- In real world, audio and visual data are not always correlated or temporally aligned.
- **Goal –** recognize and temporally localize the occurred audio or visual events in a video
- **Challenge –** weak video-level labels (lack of events' temporal and modal information) available only during training

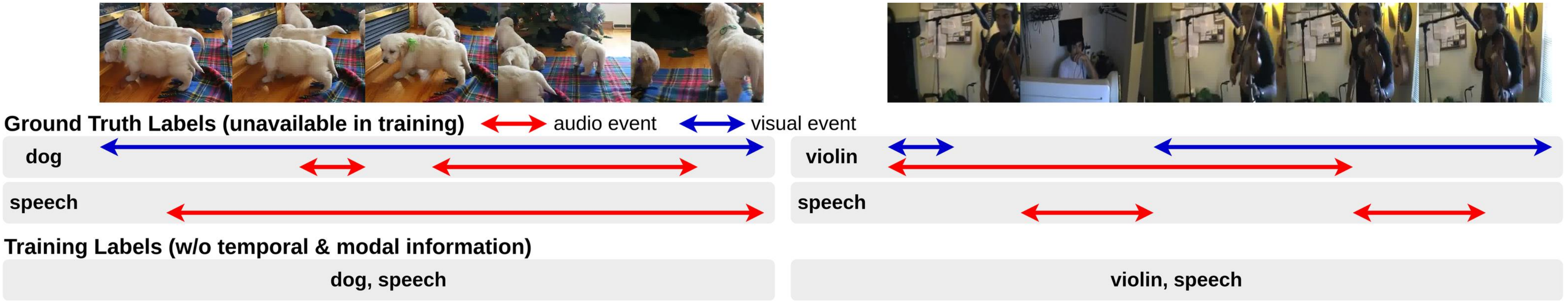


AVVP example 1

AVVP example 2

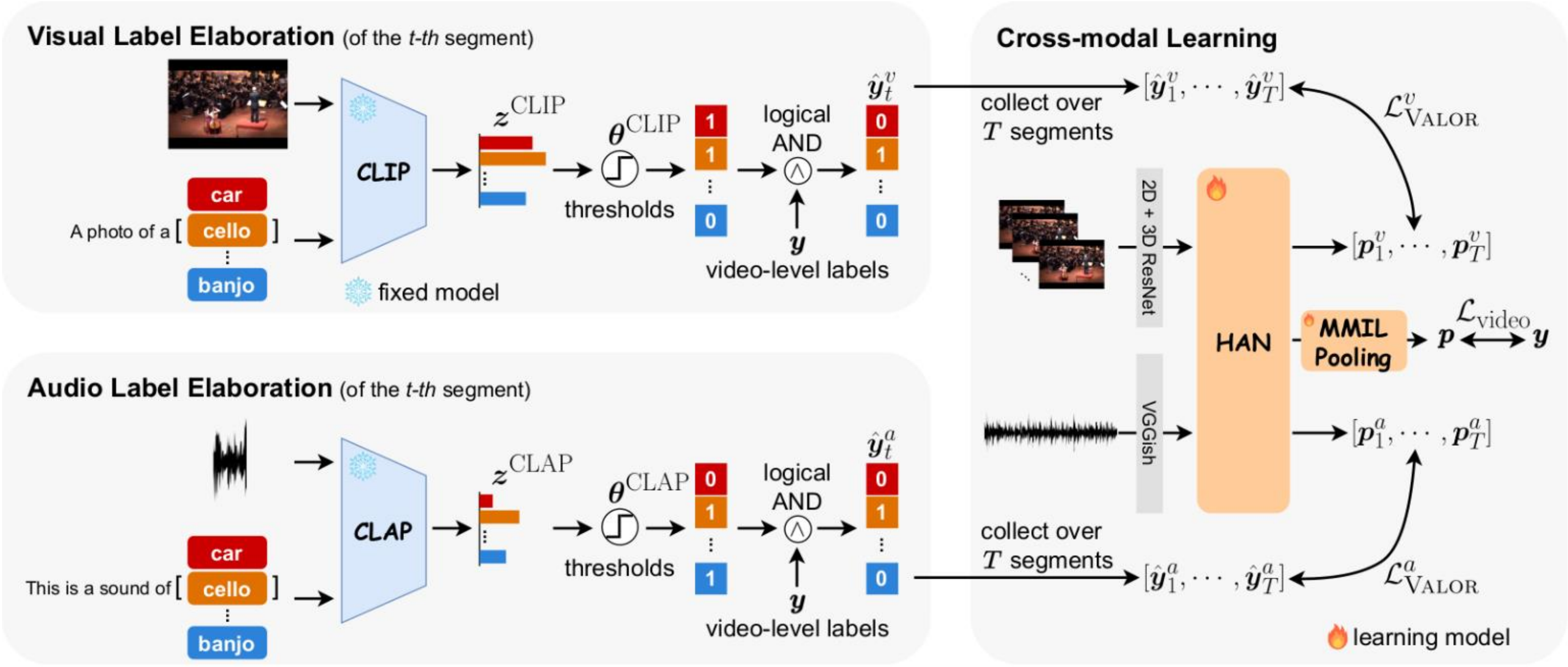
Challenges & Solutions

1. Modality independence of events' occurrence
   ➡ leverage large-scale pre-trained uni-modal contrastive models
2. Reliance on Multi-modal Multiple Instance Learning (MMIL) pooling for event modality assignment
   ➡ reliable modality-specific event labels
3. Demand for dense temporal predictions without temporal guidance during training
   ➡ faithful segment-level event labels

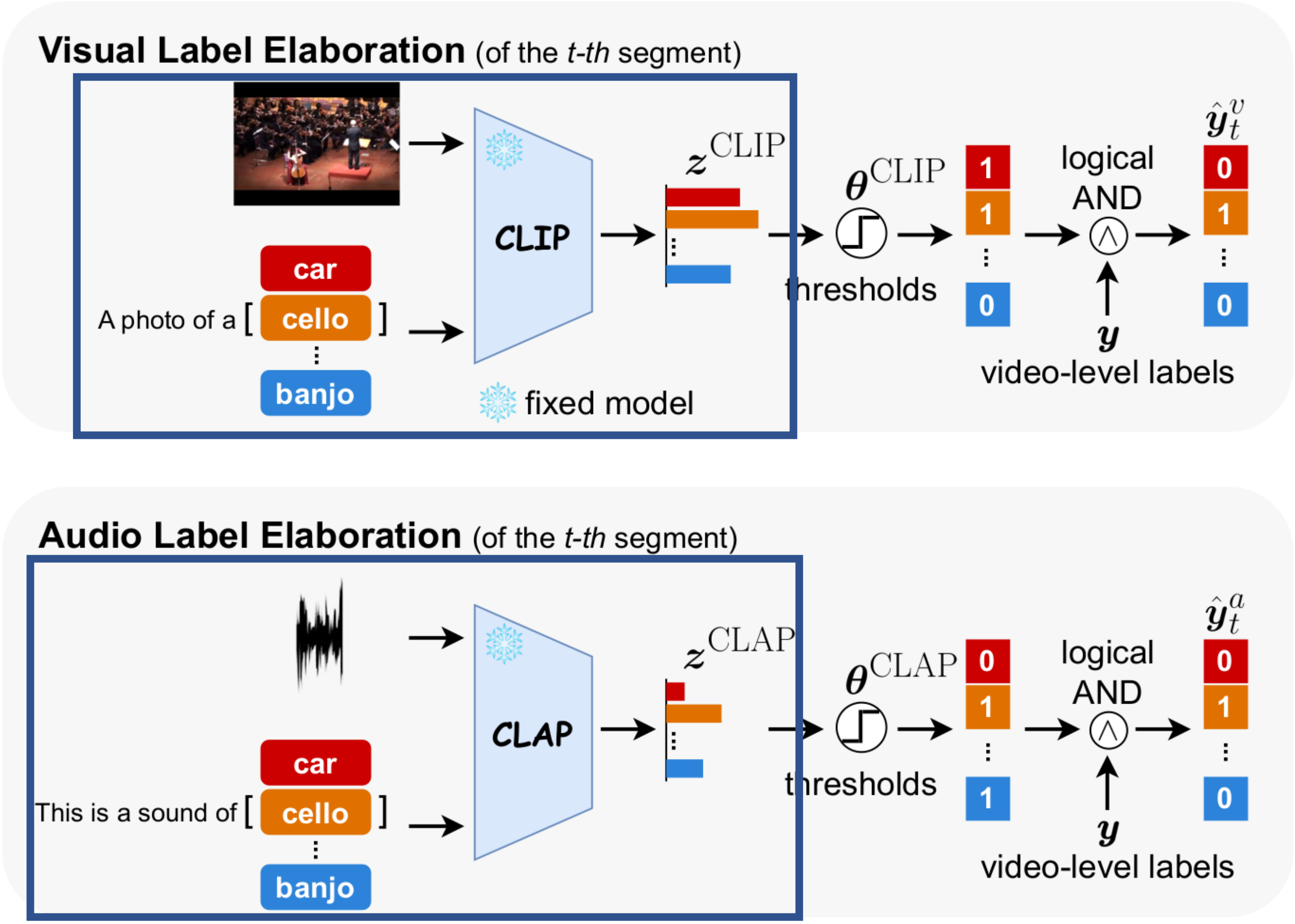


Method – Visual-Audio Label Elaboration (VALOR)

We leverage large-scale pre-trained contrastive models, CLIP and CLAP, to extract modality-aware and temporally dense training signals, $\hat{y}_t^v$ and $\hat{y}_t^a$.

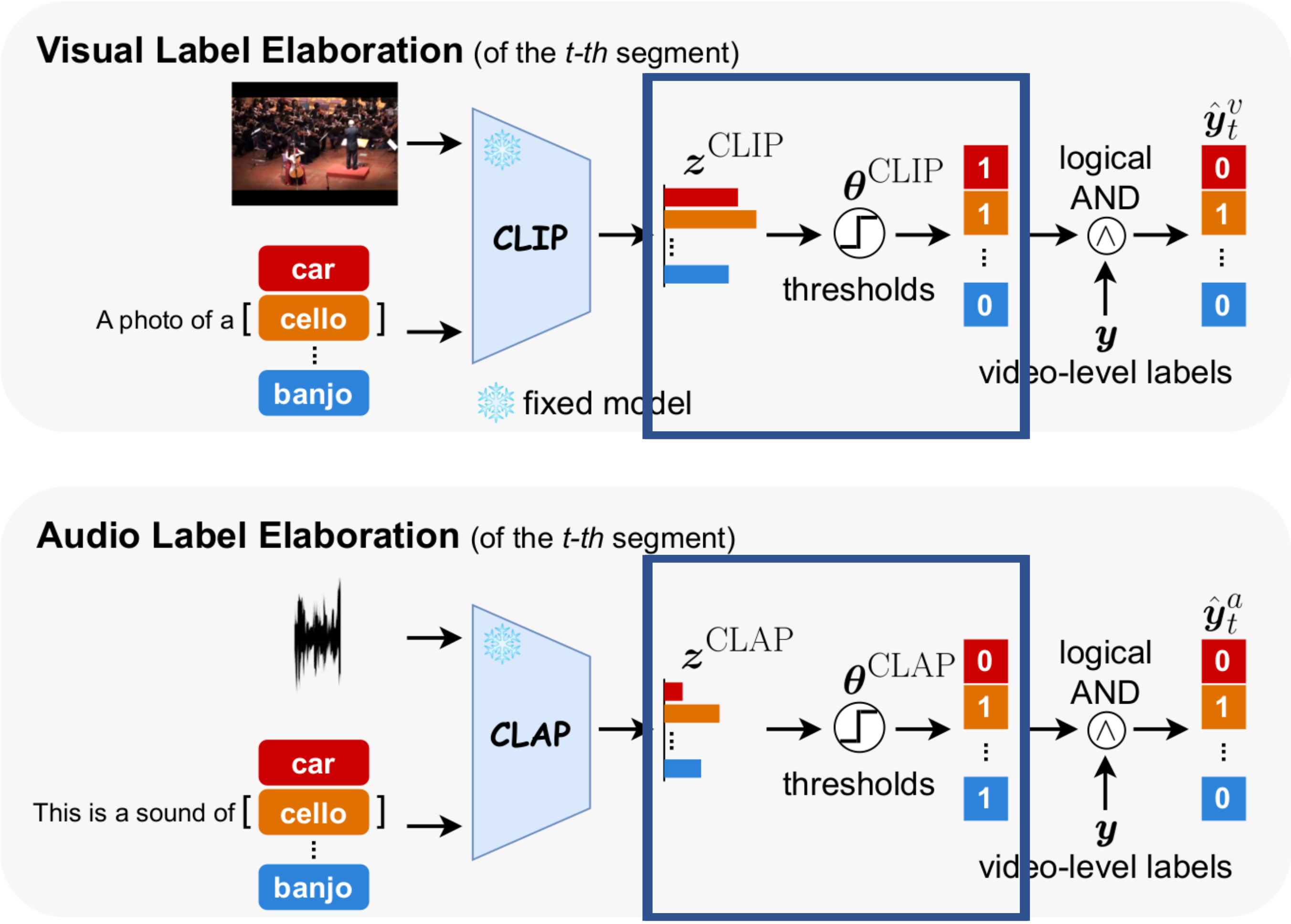


Method – Generating Modality-Specific Labels



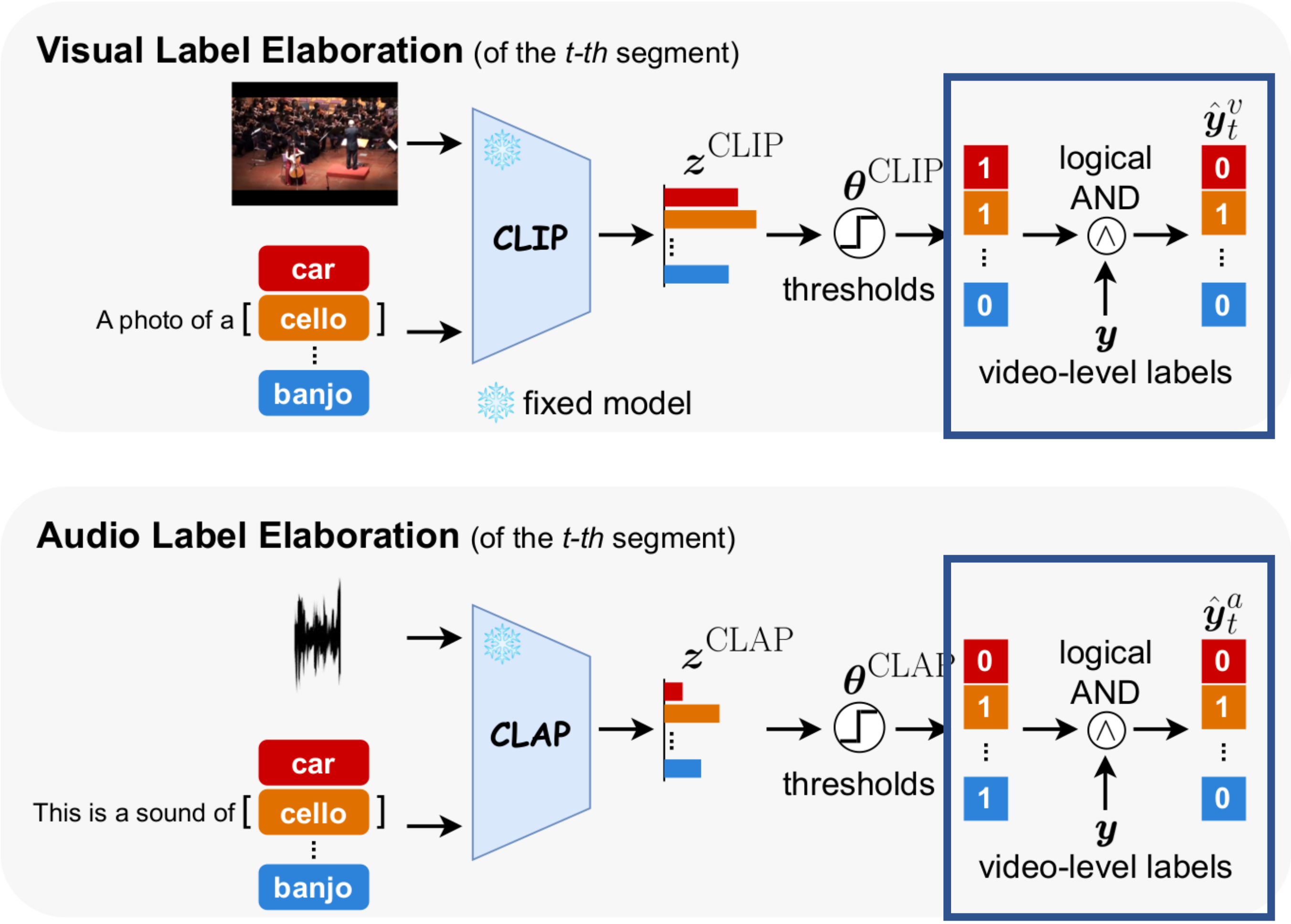
1. Generate event confidence scores z^{CLIP} and z^{CLAP} for the *t-th* segment

Method – Generating Modality-Specific Labels



1. Generate the event confidence scores z^{CLIP} and z^{CLAP} for the *t-th* segment
2. Construct segment-level labels by comparing z^{CLIP} and z^{CLAP} with the pre-defined thresholds θ^{CLIP} and θ^{CLAP}, respectively

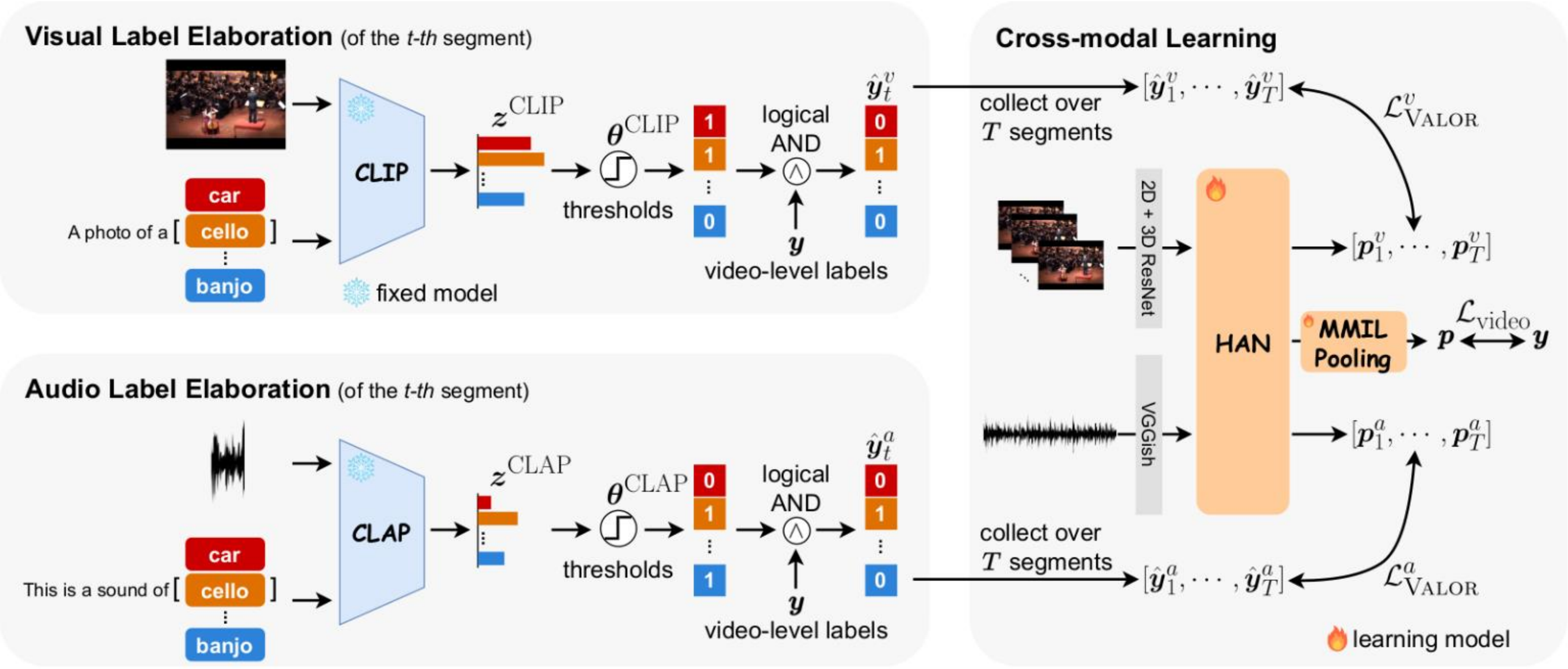
Method – Generating Modality-Specific Labels



1. Generate the event confidence scores z^{CLIP} and z^{CLAP} for the *t-th* segment
2. Construct segment-level labels by comparing z^{CLIP} and z^{CLAP} with the pre-defined thresholds θ^{CLIP} and θ^{CLAP}, respectively
3. Filter out impossible events with the given video-level labels

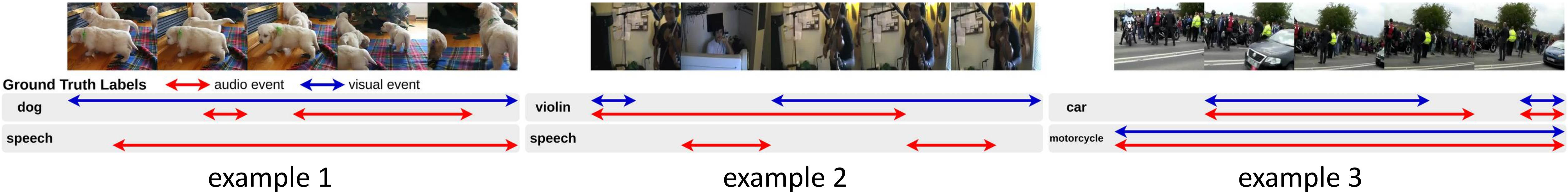
Method – Guiding Model in Cross-Modal Learning

The VALOR-generated segment-level labels in both modalities, $\hat{y}_t^v$ and $\hat{y}_t^a$, can clearly guide the model in learning where and when each event in a video occurs.



Dataset

- *Look, Listen, and Parse* (*LLP*) Dataset[1]
    - 11,849 10-second video clips
    - 25 event categories (e.g. human activities, vehicles, animals)
    - multiple events (audio, visual, or audio-visual) in a video



[1] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In ECCV, 2020. https://arxiv.org/abs/2007.10558

Quantitative Comparison – AVVP Benchmark

Methods	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
AVE [72]	47.2	37.1	35.4	39.9	41.6	40.4	34.7	31.6	35.5	36.5
AVSDN [46]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [73]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MM-Pyr [87]	60.9	54.4	50.0	55.1	57.6	52.7	51.8	44.4	49.9	50.5
MGN [51]	60.8	55.4	50.4	55.5	57.2	51.1	52.4	44.4	49.3	49.1
CVCMS [47]	59.2	59.9	53.4	57.5	58.1	51.3	55.5	46.2	51.0	49.7
DHHN [33]	61.3	58.3	52.9	57.5	58.1	54.0	55.1	47.3	51.5	51.5
MA [77]	60.3	60.0	55.1	58.9	57.9	53.6	56.4	49.0	53.0	50.6
JoMoLD [11]	61.3	63.8	57.2	60.8	59.9	53.9	59.9	49.6	54.5	52.5
VPLAN† [96]	60.5	64.8	58.3	61.2	59.4	51.4	61.5	51.2	54.7	50.8
VALOR	61.8	65.9	58.4	62.0	61.5	55.4	62.6	52.2	56.7	54.2
VALOR+	<u>62.8</u>	<u>66.7</u>	<u>60.0</u>	<u>63.2</u>	<u>62.3</u>	<u>57.1</u>	<u>63.9</u>	<u>54.4</u>	<u>58.5</u>	<u>55.9</u>
VALOR++	**68.1**	**68.4**	**61.9**	**66.2**	**66.8**	**61.2**	**64.7**	**55.5**	**60.4**	**59.0**

VALOR+: 256-dim 4-layer HAN model VALOR++: using CLAP & CLIP features

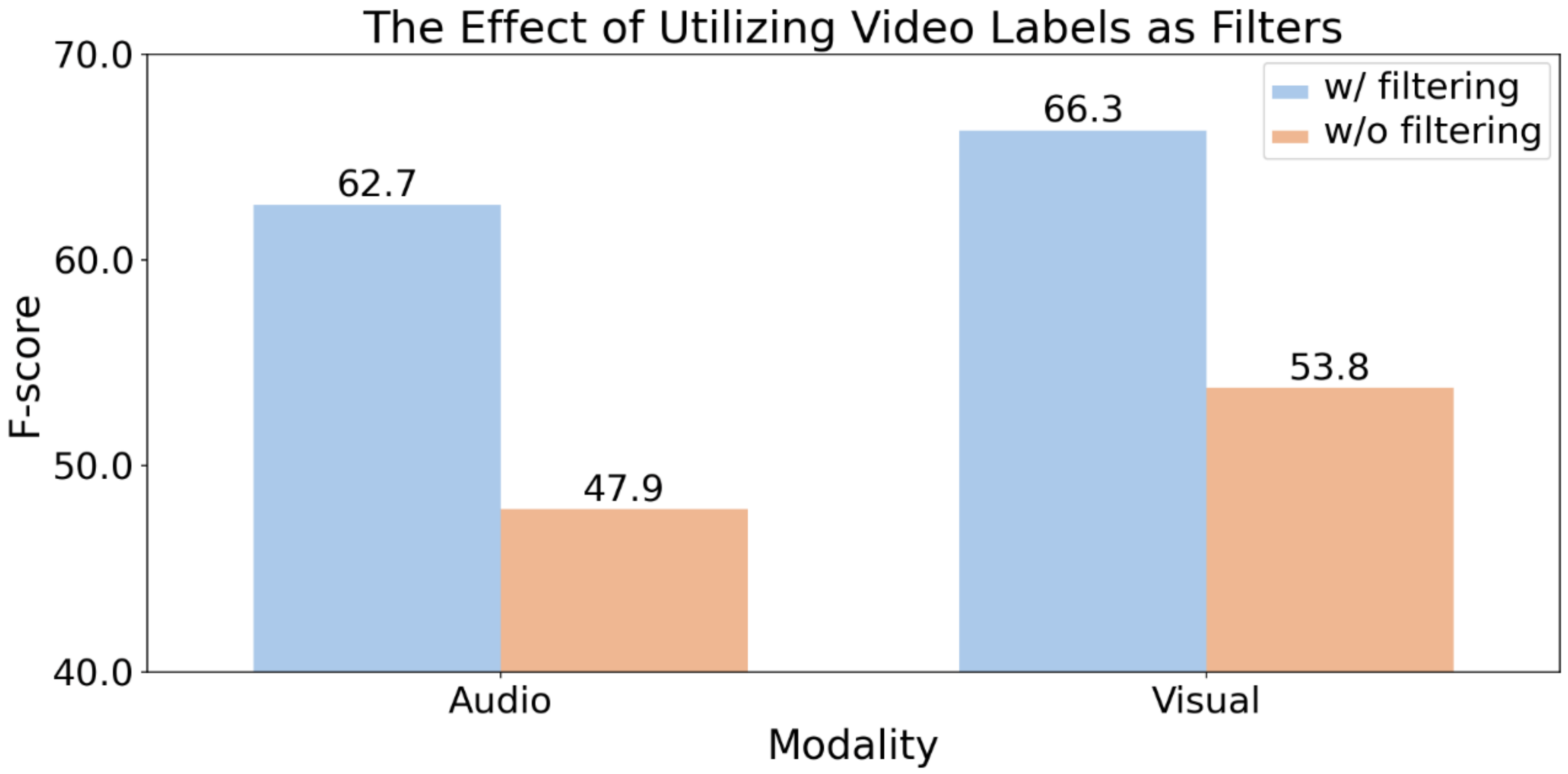
Ablation Study – How to Choose the Labeler

- We demonstrate the necessity and importance of **using large-scale pre-trained uni-modal models** to annotate **modality-aware segment-level labels**.

Dense Labeler	Modality Label	Segment-level A	Segment-level V	Segment-level AV	Segment-level Type	Segment-level Event	Event-level A	Event-level V	Event-level AV	Event-level Type	Event-level Event
None	✔	62.0	54.5	50.2	55.6	57.1	53.5	50.5	43.6	49.2	50.3
HAN	✔	62.1	56.4	52.1	56.8	57.6	53.4	52.0	45.4	50.3	50.6
CLIP&CLAP	✘	41.0	59.0	34.5	44.9	52.1	33.2	56.2	28.2	39.2	43.1
CLIP&CLAP	✔	**62.7**	**66.3**	**61.0**	**63.4**	**61.8**	**55.5**	**62.0**	**54.1**	**57.2**	**53.8**

Ablation Study – Whether Using Video Labels as Filters

- We employ video-level labels to eliminate impossible events misclassified by CLIP or CLAP for generating reliable pseudo labels.



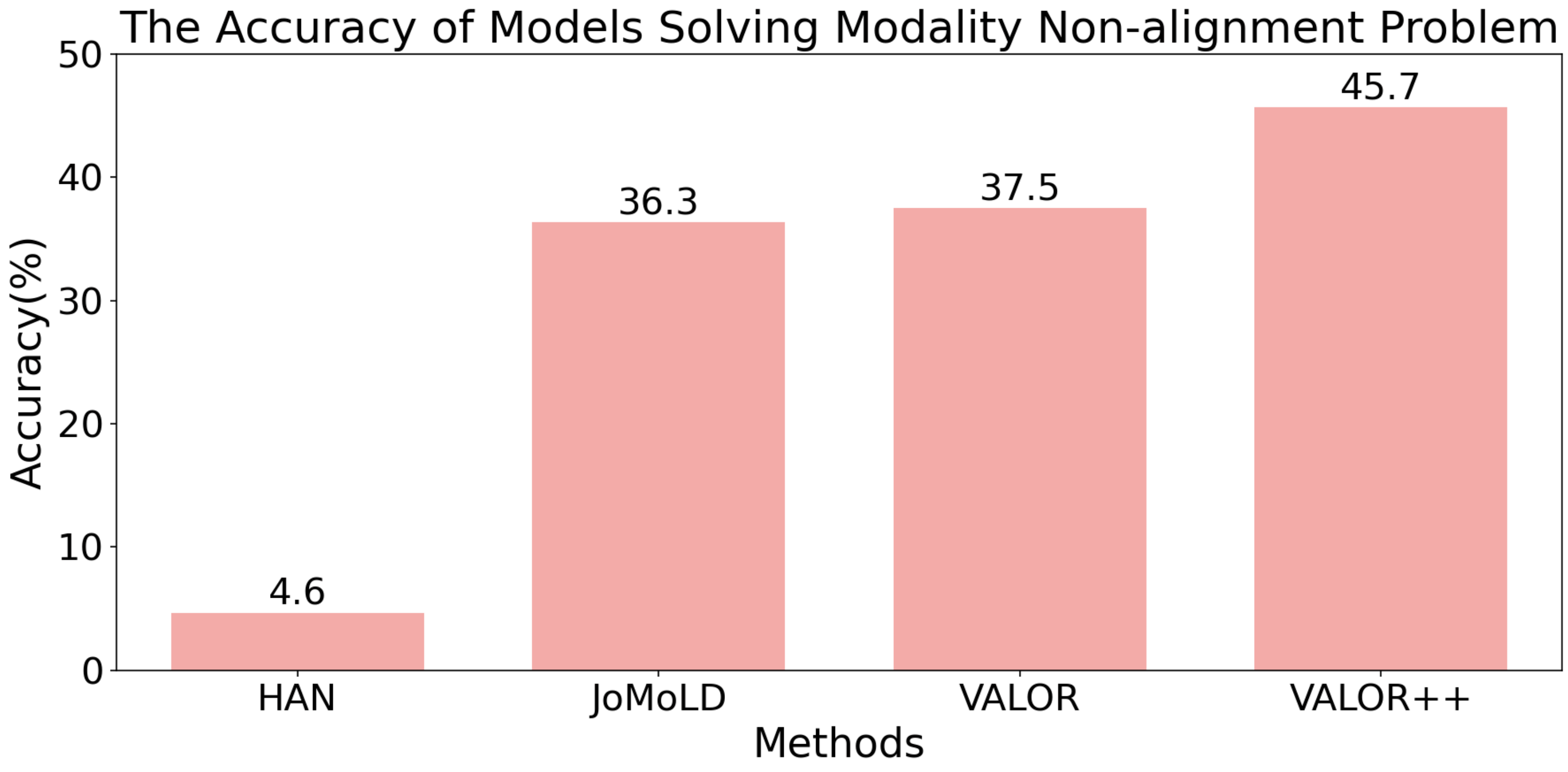
Ablation Study – How Accurate Are the Elaborated Labels

- We compare VALOR to a naive approach where we assume video-level labels also serve as segment-level labels.

Label Generation Methods	Audio	Visual	Audio-Visual
Video Labels	80.08	67.21	59.45
VALOR	**85.07** (+4.99)	**82.14** (+14.93)	**77.07** (+17.62)

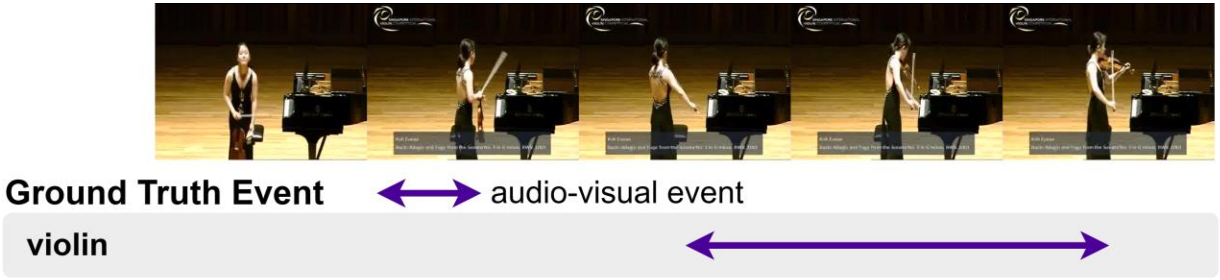
Ablation Study – Address the Modality Non-alignment Problem

- We assess how well the models can correctly predict the modality non-aligned events.
  - 4048 segment-level events are modality non-aligned (occurring in exactly one modality)



Quantitative Comparison – Generalizability of VALOR

- We showcase the generalizability of VALOR by applying it to the Audio-Visual Event Localization (AVE) task.
- Audio-Visual Event Localization
  - One video only contains one audio-visual event.
  - A video is labeled as the event if the event is audible and visible in the segment.



Method	Accuracy(%)
VGG-like, VGG-19 features	
AVEL [72]	66.7
AVSDN [46]	67.3
CMAN [85]	70.4
AVRB [58]	68.9
AVIN [57]	69.4
AVT [44]	70.2
CMRAN [82]	72.9
PSP [95]	73.5
CMBS [80]	74.2
VGG-like, Res-151 features	
AVEL [72]	71.6
AVSDN [46]	74.2
CMRAN [82]	75.3
CMBS [80]	76.0
CLAP, CLIP, R(2+1)D features	
HAN	75.3
VALOR	**80.4**

Thanks