# **Empowering LLMs with Pseudo-Untrimmed Videos for Audio-Visual Temporal Understanding**

Stage I

Stage II

Template

Pseudo Annotations

Audio Adapter

15s – 21s: A train stopped at the tunnel ntrance to make a hum of power, and ther

21s - 40s: A train raced along the tracks,

40s - 55s: The train crept into the station, and the whistle of the locomotive hissed.

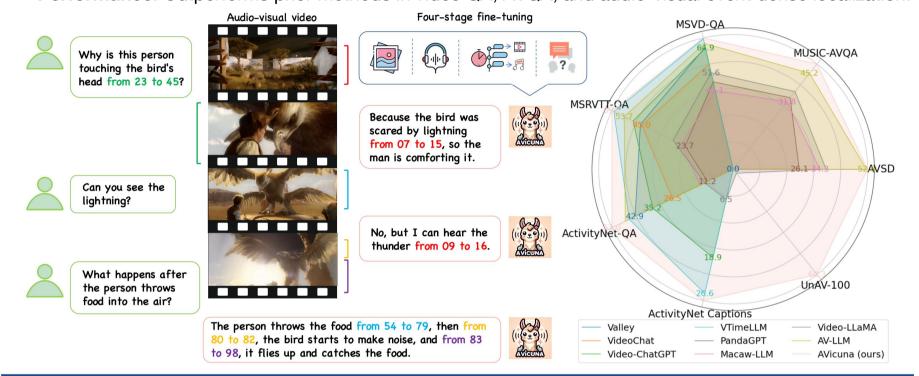
Yunlong  $Tang^1$ ,  $Daiki Shimada^2$ ,  $Jing Bi^1$ ,  $Minggian Feng^1$  $Hang\ Hua^1$ , Chenliang  $Xu^1$ 

<sup>1</sup>University of Rochester, <sup>2</sup>Sony Group Corporation



#### Contribution

- **PU-VALOR Dataset**: Proposes a method (event-based clustering + temporal scaling and permutation) to generate 114K+ pseudo-untrimmed videos from trimmed clips, enabling scalable training with precise temporal annotations.
- **AVicuna Model**: Integrates an Audio-Visual Token Interleaver (AVTI) into an audio-visual LLM to align multimodal tokens, supporting event localization and temporal dialogue.
- **A5-222K Dataset**: Curates 222K audio-text pairs to strengthen audio-text alignment during training.
- Performance: Outperforms prior methods in video QA, AVQA, and audio-visual event dense localization.



#### Motivation

- Missing fine-grained temporal annotations in untrimmed data: Existing datasets (e.g., VALOR-32K) lack precise event-time labels, limiting models' ability to align audio-visual events with text.
- Inefficient multimodal-temporal modeling: Prior methods (e.g., Video-LLaMA) fail to synchronize audio-visual cues over time.

Video–Text Pair

Dataset	Un- trimmed	Audio- Visual	Captions	Time- stamps	
ActNetCaps (Krishna et al. 2017)	<b> </b>	×	<b>√</b>	<b>√</b>	
InternVid (Wang et al. 2023d)	✓	×	$\checkmark$	$\checkmark$	
VGG-Sound-AVEL100K	×	$\checkmark$	×	$\checkmark$	
AVVP (Tian, Li, and Xu 2020)	×	$\checkmark$	×	$\checkmark$	
LFAV (Hou et al. 2023)	✓	$\checkmark$	×	$\checkmark$	
UnAV-100 (Geng et al. 2023)	✓	$\checkmark$	×	$\checkmark$	
VALOR (Chen et al. 2023b)	×	$\checkmark$	$\checkmark$	×	
PU-VALOR (ours)	<b>√</b>	$\checkmark$	$\checkmark$	$\checkmark$	

#### Clustering PU-VALOR video\_2 video\_k VALOR-32K Dataset Dataset • • • caption\_2 caption\_k caption\_1 Random Temporal Scaling & Permutation **Embedding** "event": "At which tervals in the video can identify {caption\_i}?", Text Encoder "timestamps": [ {pseudo start\_time}, 10s Short Video {pseudo end\_time} Audio-visual "sub\_video\_id": {video\_i Video Captions

Pseudo-Untrimmed Video Dataset Curation

Pipeline for creating the PU-VALOR dataset, which involves extracting text embeddings from high-quality audiovisual captions of the original trimmed VALOR-32K dataset, clustering these embeddings, and then applying Random Temporal Scaling & Permutation to generate pseudo-untrimmed videos.

Pseudo-Untrimmed Videos ( ): Pseudo Start/End Timestamps)

### Pseudo-Untrimmed Video Dataset Curation (Continue)

- 1. Source: Trimmed Clips: Use existing datasets (e.g., VALOR-32K) with 10s clips and captions as base material 2. **Event-Based Clustering**: Group clips by semantic events using caption keywords.
- 3. **Temporal Scaling**: Adjust clip speed to simulate duration variations while preserving A/V sync.
- 4. **Permutation**: Shuffle scaled clips within clusters and concatenate to create long pseudo-untrimmed videos. 5. Annotation Transfer: Propagate original timestamps to synthetic videos with scaling/permutation offsets.

?

User's query

**(**ֆոի ֆ

6. Quality Control: Filter incoherent sequences and add background noise for realism.

OUTPUT: PU-VALOR (114K+ videos): Fine-grained temporal labels (event start/end times with diverse event transitions (e.g., "pour water  $\rightarrow$  boil  $\rightarrow$  stir").

#### AVicuna Model

#### Stage III & IV \_\_\_\_\_ Vision Caption We can observe <event> from <start\_time> to <end\_time>, in 1 Aligned Response both imagery and sound. Vicuna 1 ♦ LoRA Vicuna Visual Adapter 4 1 Describe the visual content. Vision Encoder At which intervals in Audio-Visual Tokens Interleaver the video can we identify **Audio Caption** <event>, either 1 Audio Adapter Visual Adapter by watching or listening? Vicuna Vision Encoder Audio Encoder 1

**AVicuna Model Architecture**. Vision and Audio Adapters are MLPs that align modalities with LLM. The Audio-Visual Tokens Interleaver ensures temporal synchronization. LoRA fine-tuning aligns temporal boundaries with events and enhances instruction-following capabilities.

Video

Describe the

audio content.

#### Four-Stage Training

- Stage I: Vision-Text Alignment: Freeze the LLM and update only the Vision Adapter using image-text pairs (LCS-558K) to align visual features with the LLM's token space.
- Stage II: Audio-Text Alignment: Freeze the LLM and update only the Audio Adapter using audio-text pairs (A5-222K) to map audio features into the LLM's
- Stage III: Time-Event Alignment: Freeze both adapters and fine-tune the connective adapter (via LoRA) in the LLM with temporally annotated QA pairs (PU-VALOR + InternVid) to link temporal boundaries with multimodal events.
- Stage IV: Instruction Tuning: Fine-tune the entire model on diverse instruction datasets (e.g., UnAV-100, VideoInstruct100K, ActivityNet Captions, DiDeMo) to boost general query following and recover QA performance

Comparison with existing LLM-based

QA, MSRVTT-QA, ActivityNetQA) and

methods on open-ended video QA (MSVD-

AVQA (AVSD, MUSIC-AVQA) benchmarks.

## **AAAI-25 / IAAI-25 / EAAI-25**

FEBRUARY 25 - MARCH 4, 2025 | PHILADELPHIA, USA

**Experiments** (Continue)



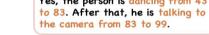




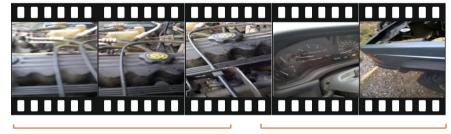








Predict temporal intervals for the audio-visual events





When can you find engine knocking?

We can find engine knocking from 00 to

the soundtrack.





What can you see or hear

We can observe people eating from 09 to 99, either through sight or sound in the video.

Reasoning given audio-visual context





How many instruments in the video did not sound from beginning to end?

There are no instruments in the video that did not sound from beginning to end.





Visit our GitHub repo for code <u> https://github.com/yunlong10/AVicuna</u>

Acknowledgement: This work was supported by Sony Group Corporation. We would like to thank Sayaka Nakamura and Jerry Jun Yokono for insightful discussion.

## **Experiments**

Untrimmed or

Pseudo-Untrimmed

Method	A&V	TU	#Pairs	LLM-size	AVSD	MUSIC-QA	MSVD-QA	MSRVTT-QA	ActivityNet-QA
Valley (Luo et al. 2023b)	x	×	1.5M	13B	_	-	65.4	45.7	26.5
VideoChat (Li et al. 2023b)	×	$\checkmark$	25M	7B	-	-	56.3	45.0	26.5
Video-ChatGPT (Maaz et al. 2023)	×	$\checkmark$	0.9M	7B	_	-	64.9	49.3	35.2
VTimeLLM (Huang et al. 2023)	×	$\checkmark$	0.7M	7B	_	-	69.8	58.8	45.5
PandaGPT (Su et al. 2023)	✓	×	128M	13B	26.1	33.7	46.7	23.7	11.2
Macaw-LLM (Lyu et al. 2023a)	✓	×	0.3M	7B	34.3	31.8	42.1	25.5	14.5
AV-LLM (Shu et al. 2023)	✓	×	1.6M	13B	52.6	45.2	67.3	53.7	47.2
Video-LLaMA (Zhang et al. 2023b)	✓	$\checkmark$	2.8M	7B	36.7	36.6	51.6	29.6	12.4
AVicuna (ours)	✓	$\checkmark$	1.1M	7B	53.1	49.6	70.2	59.7	53.0

Method	0.5	0.6	0.7	0.8	0.9	Avg.
VSGN (Zhao et al. 2021)	24.5	20.2	15.9	11.4	6.8	24.1
TadTR (Liu et al. 2022)	30.4	27.1	23.3	19.4	14.3	29.4
ActionFormer (Zhang et al. 2022a)	43.5	39.4	33.4	27.3	17.9	42.2
UnAV (Geng et al. 2023)	50.6	45.8	39.8	32.4	21.1	47.8
UniAV-AT (Geng et al. 2024)	54.1	48.6	42.1	34.3	20.5	50.7
UniAV-ST (Geng et al. 2024)	54.8	49.4	43.2	35.3	22.5	51.7
AVicuna (ours)	60.0	50.4	49.6	43.5	36.5	60.3

Comparison of the results on the UnAV-100 dataset for			
the AVEDL task.	1	com	

0.9 Avg. Setting 0.6 0.7 0.8 **AVicuna** w/o PU-VALOR 19.5 14.3 10.2 6.8 4.5 27.9 50.1 45.2 40.2 34.2 29.4 51.1 w/o AVTI 22.2 16.5 11.4 6.8 2.7 30.1 w/o A5-222K w/o Audio 29.0 23.9 18.8 13.6 8.8 35.8

Ablation study on the dataset and model mponents, which lead to decreases in mAP.

mAP Scores at Different Audio-Interleaving Rates • → mAP@0.9

AVicuna's performances on UnAV-100 measured by mAP scores at different AIRs (Audio-Interleaving Rates).

--**+**-- Avg. mAP