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

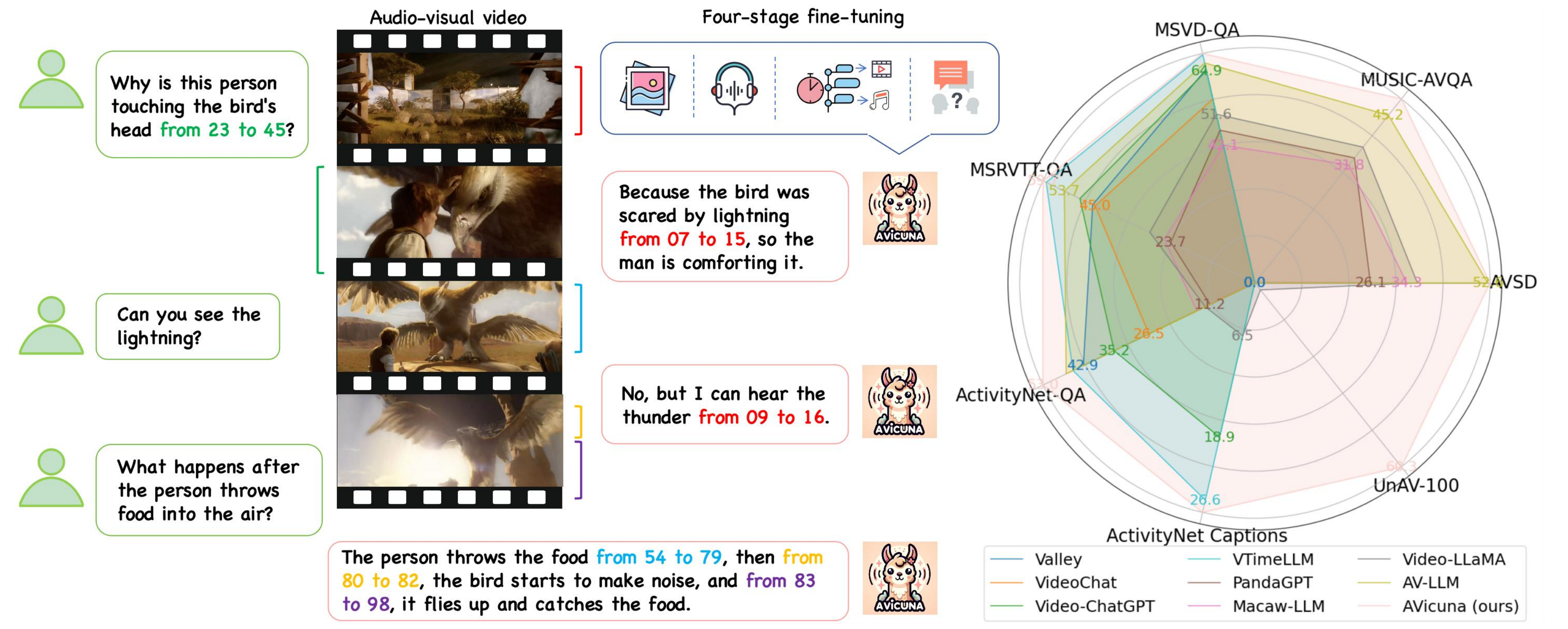
Yunlong Tang[1], Daiki Shimada[2], Jing Bi[1], Mingqian Feng[1], Hang Hua[1], Chenliang Xu[1]

[1]University of Rochester, [2]Sony Group Corporation

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

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.

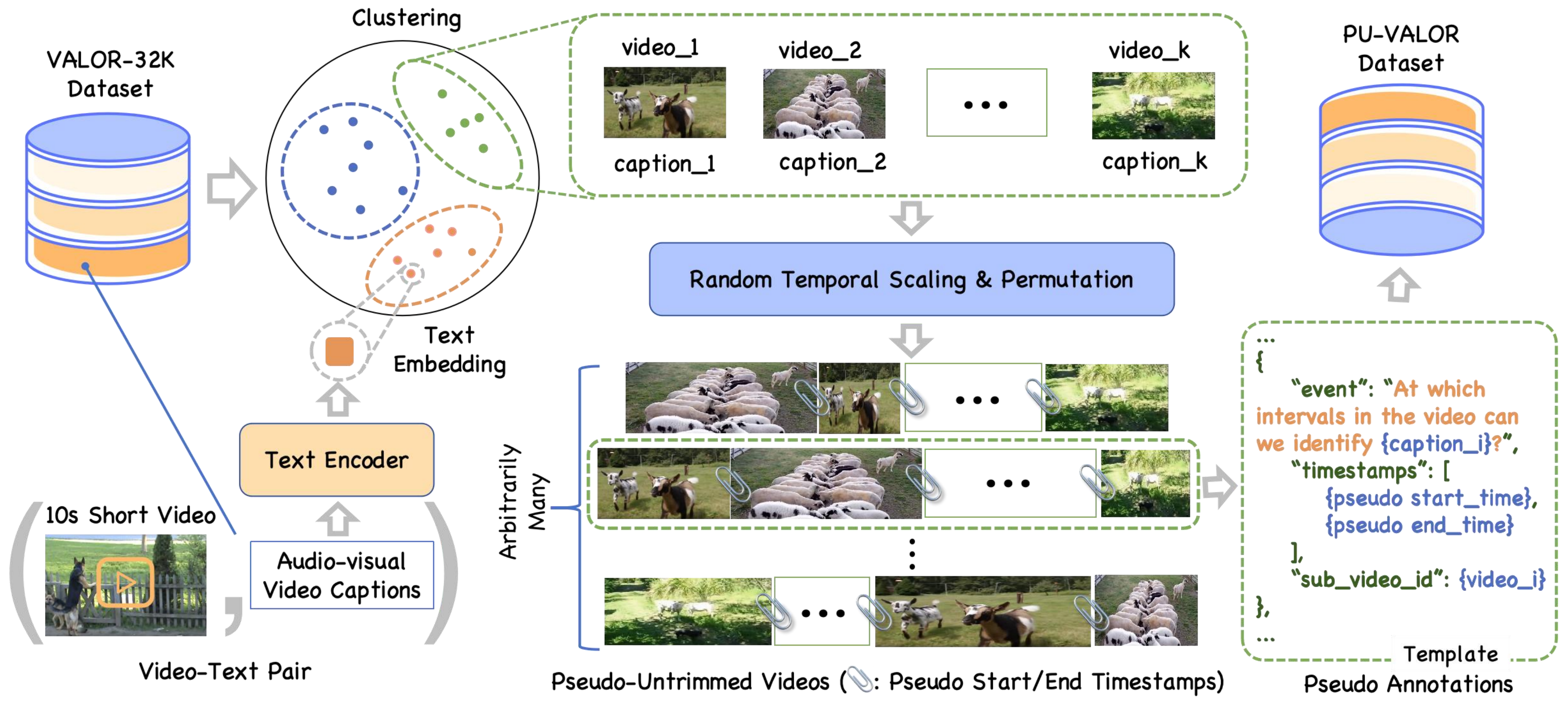


Motivation

- 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.

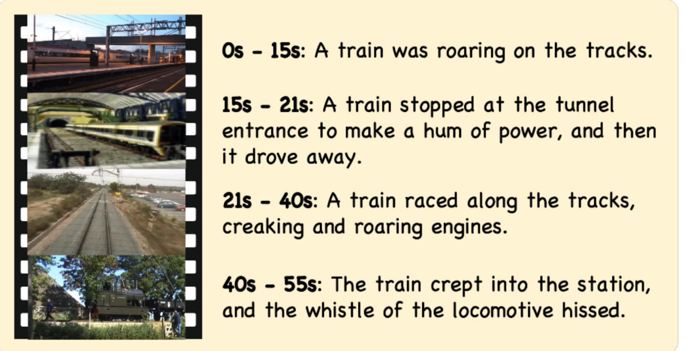
Dataset	Un-trimmed	Audio-Visual	Captions	Time-stamps
ActNetCaps (Krishna et al. 2017)	✓	×	✓	✓
InternVid (Wang et al. 2023d)	✓	×	✓	✓
VGG-Sound-AVEL100K	×	✓	×	✓
AVVP (Tian, Li, and Xu 2020)	×	✓	×	✓
LFAV (Hou et al. 2023)	✓	✓	×	✓
UnAV-100 (Geng et al. 2023)	✓	✓	×	✓
VALOR (Chen et al. 2023b)	×	✓	✓	×
PU-VALOR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

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.

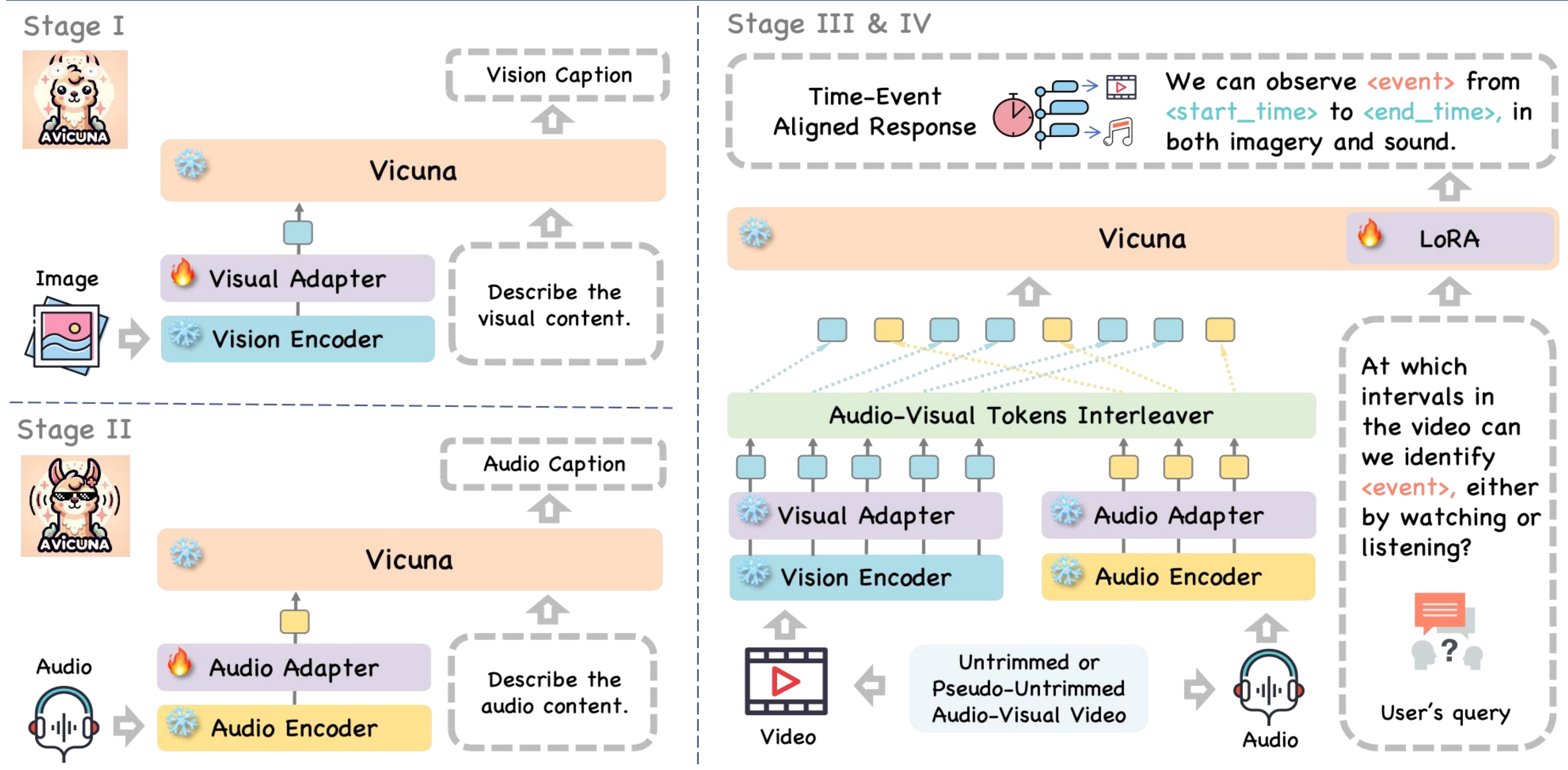
Pseudo-Untrimmed Video Dataset Curation (Continue)



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.
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 → boil → stir").

AVicuna Model



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.

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 space.
- *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.

Experiments

Method	A&V	TU	#Pairs	LLM-size	AVSD	MUSIC-QA	MSVD-QA	MSRVTT-QA	ActivityNet-QA
Valley (Luo et al. 2023b)	×	×	1.5M	13B	-	-	65.4	45.7	26.5
VideoChat (Li et al. 2023b)	×	✓	25M	7B	-	-	56.3	45.0	26.5
Video-ChatGPT (Maaz et al. 2023)	×	✓	0.9M	7B	-	-	64.9	49.3	35.2
VTimeLLM (Huang et al. 2023)	×	✓	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)	✓	✓	2.8M	7B	36.7	36.6	51.6	29.6	12.4
AVicuna (ours)	✓	✓	1.1M	7B	**53.1**	**49.6**	**70.2**	**59.7**	**53.0**

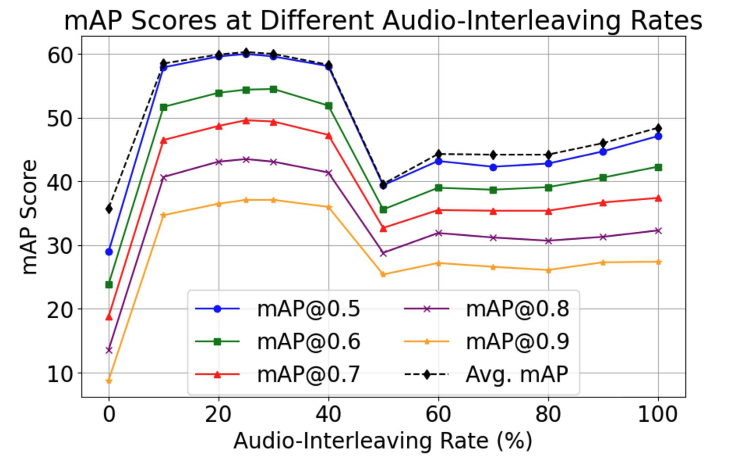
Comparison with existing LLM-based methods on open-ended video QA (MSVD-QA, MSRVTT-QA, ActivityNetQA) and AVQA (AVSD, MUSIC-AVQA) benchmarks.

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.

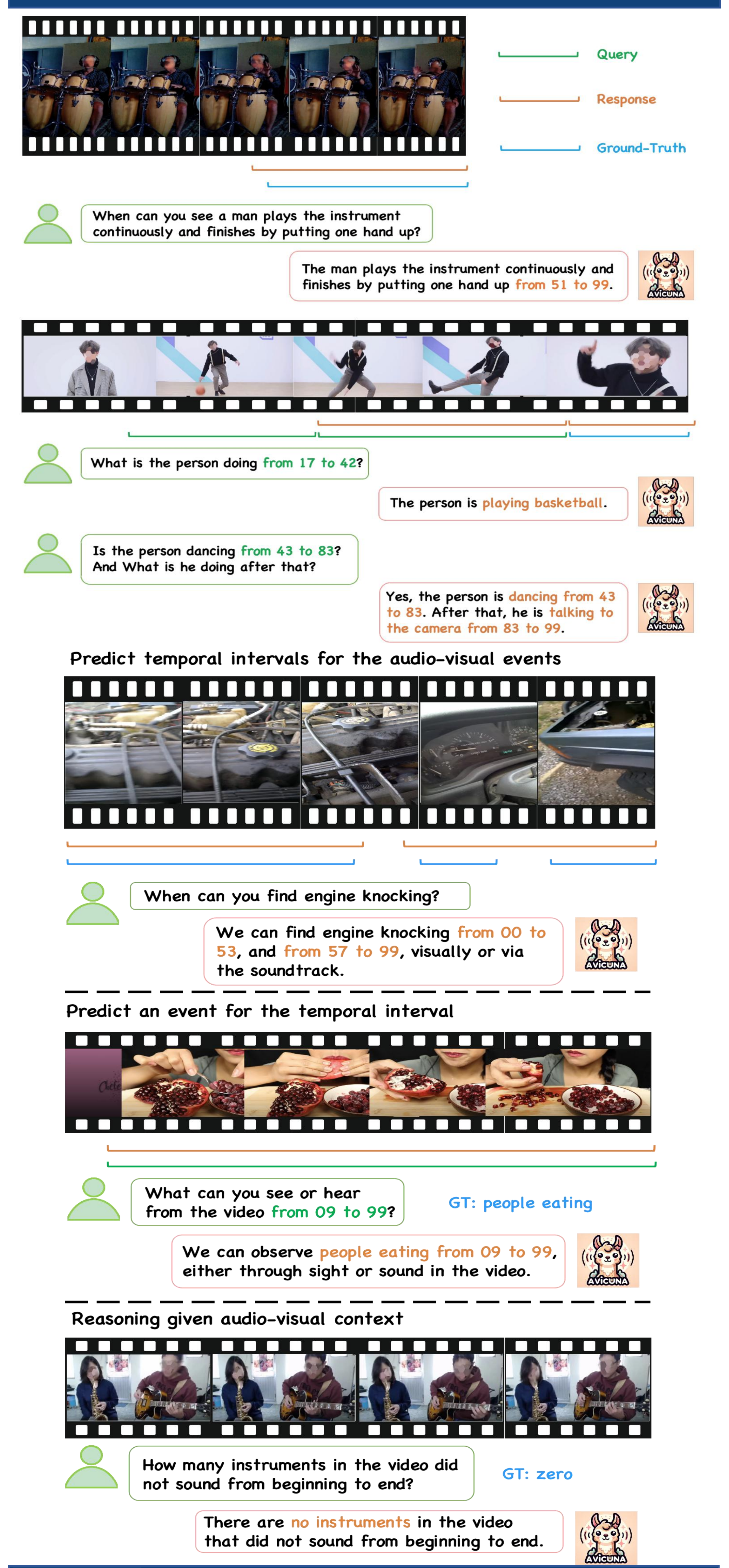
Setting	0.5	0.6	0.7	0.8	0.9	Avg.
AVicuna	**60.0**	**54.4**	**49.6**	**43.5**	**37.1**	**60.3**
w/o PU-VALOR	19.5	14.3	10.2	6.8	4.5	27.9
w/o AVTI	50.1	45.2	40.2	34.2	29.4	51.1
w/o A5-222K	22.2	16.5	11.4	6.8	2.7	30.1
w/o Audio	29.0	23.9	18.8	13.6	8.8	35.8

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



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

Experiments (Continue)



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

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