Towards Open-Vocabulary Audio-Visual Event Localization
Supplementary Material

Jinxing Zhou[1] Dan Guo[2] Ruohao Guo[3] Yuxin Mao[4] Jingjing Hu[2]
Yiran Zhong[5] Xiaojun Chang[1,6] Meng Wang[2,*]
[1]MBZUAI [2]HFUT [3]PKU [4]NWPU [5]OpenNLPLab [6]USTC
https://github.com/jasongief/OV-AVEL

Table A1. **Ablation study on the number of temporal layers L.** Results are reported on the total test data.

L	Acc.	Seg.	Eve.	*Avg.*
1	**67.1**	**56.9**	**49.5**	**57.8**
2	65.4	56.0	49.2	56.9
3	62.8	54.0	47.3	54.7

Table A2. **Ablation study on the employment of the text** `other`. 'TF' and 'FT' represent the training-free baseline and fine-tuning baseline, respectively.

	Data type	w. `other`				w. `background`			
		Acc.	Seg.	Eve.	*Avg.*	Acc.	Seg.	Eve.	*Avg.*
TF	total	59.2	46.7	34.0	**46.6**	59.1	46.6	33.8	46.5
	seen	57.5	45.0	34.0	**45.5**	57.5	45.1	34.0	**45.5**
	unseen	59.8	47.3	34.0	**47.0**	59.7	47.2	33.7	46.9
FT	total	67.1	56.9	49.5	**57.8**	66.2	56.1	48.5	56.9
	seen	72.5	61.8	54.5	**62.9**	71.8	60.9	53.7	62.1
	unseen	64.9	55.0	47.5	**55.8**	63.9	54.1	46.4	54.8

A. The number of temporal layers L

Our fine-tuning baseline employs L learnable temporal layers to enhance temporal interactions within audio and visual modalities. The results, as shown in Table A1, illustrate the impacts of varying the number of layers. The model achieves the highest average performance using only one temporal layer. Increasing the number of temporal layers may make the model more complex and lead to overfitting, thus degrading the performance. Consequently, we identify $L = 1$ to implement the temporal layer, which is lightweight and only introduces 8.4M trainable parameters.

Table A3. **Temporal interactions in intra- and cross- modalities for model fine-tuning.** Results are reported on the total test data.

Cases	Acc.	Seg.	Eve.	*Avg.*
intra only	**67.1**	**56.9**	**49.5**	**57.8**
cross only	54.4	45.9	39.2	46.5
intra + cross	63.5	54.3	47.1	55.0

B. Further Ablation Study on `other`

In Sec. 4.3 of our main paper, we have demonstrated that our baseline models using additional class text `other` outperform models that do not use `other`. Here, we further compare the employment of `other` with another option, namely `background`. The experimental results are shown in Table A2. For both the training-free and fine-tuning baselines, the use of `other` is slightly better than `background`. Compared to `background`, we speculate that the text `other` can further help the model deal with situations that include other meaningful event classes not listed in the seen and unseen class texts.

C. Intra-modal vs. Cross-modal temporal layers

The temporal layers in our fine-tuning baseline facilitate temporal interactions within the audio and visual modalities (*intra-modal*). We also attempted to insert some temporal layers to capture *cross-modal* temporal relations. As shown in Table A3, adding cross-modal temporal layers does not yield improvements. We speculate that the audio and visual features extracted by the pretrained ImageBind model can provide explicit and precise semantics of audio events and visual events, reducing the need for cross-modal interactions. By focusing on the temporal interactions in intra-modality, the model can achieve satisfactory performance.

D. Comparison with the CLIP&CLAP

In sec. 4.2, we compare our training-free baseline with another zero-shot approach, CLIP&CLAP. Here, we provide

Table A4. **Comparison between the Training-free baseline with the variant CLIP&CLAP.** The default implementation in our main paper uses ImageBind [4] to *jointly* extract multimodal features and generate audio-visual event predictions. In contrast, the *separate* variant uses the pretrained CLAP [8] and CLIP [6] models to extract features independently and computes the audio-text and visual-text feature similarities separately.

Data type	ImageBind (*joint*)				CLAP&CLIP (*separate*)			
	Acc.	Seg.	Eve.	*Avg.*	Acc.	Seg.	Eve.	*Avg.*
total	59.2	46.7	34.0	**46.6**	51.5	41.9	31.7	41.7
seen	57.5	45.0	34.0	**45.5**	51.4	41.4	31.9	41.6
unseen	59.8	47.3	34.0	**47.0**	51.6	42.2	31.6	41.8

more implementation details. The training-free baseline introduced in our main paper utilizes ImageBind [4] to extract audio, visual, and textual embeddings. It computes the audio-text and visual-text feature (cosine) similarities to determine final audio-visual event predictions. We refer to this strategy as ***joint*** since multimodal features are extracted from a shared feature space. Furthermore, we compare this approach with another variant, where the audio-text and visual-text feature similarities are calculated using feature embeddings from ***separate*** backbones. Specifically, for each segment, the pretrained CLAP [8] model is used to extract the audio and text features to generate the audio-text feature similarity; the pretrained CLIP [6] model is used to extract the visual and text features to generate the visual-text feature similarity. Notably, the text encoders of CLAP and CLIP models are different, so the text features are extracted independently. After obtaining the audio-text and visual-text feature similarities, we identify the event categories of the audio segments and visual segments based on the highest similarity values. The final audio-visual event prediction can be made by comparing the consistency of the predicted audio and visual event categories. The experimental results are shown in Table A4. The ***joint*** baseline model using ImageBind significantly outperforms the ***separate*** variant, with improvements of 4.9%, 3.9%, and 5.2% in the *Avg.* metric on the total, seen, and unseen test data, respectively. These results indicate the advantages of adopting a joint feature space for multimodal feature embedding, which can better capture semantic alignment among multiple modalities for the OV-AVEL task.

E. Zero-shot Evaluation on AVE [7] Dataset

The AVE dataset is constructed for the closed-set AVEL task [7]. Here, we directly apply our two baseline models to the test set of AVE dataset in a zero-shot inference manner. As shown in Table A5, the fine-tuning baseline continues to outperform the training-free version, demonstrating results competitive with the prior unsupervised state-of-the-art (SOTA) method CMLCL [1]. Notably, CMLCL still uses unlabeled videos of the training set in the AVE dataset. Moreover, if further fine-tuning our baseline model on the AVE dataset, the model can reach 79.6% accuracy without sophisticated designs, approaching the performance of fully-supervised AVEL methods [3, 7, 9–11]. Nevertheless, we encourage readers to focus on the intrinsic differences: our method is designed for the open-vocabulary AVEL, while prior SOTA methods are tailored specifically for closed-set AVEL.

Table A5. **Zero-shot evaluation on AVE [7] dataset.**

Manners	Methods	Acc.
zero-shot	training-free (our)	54.8
	fine-tuning (our)	61.9
unsupervised	CMLCL [1]	63.2

F. Class-wise Performance of the Proposed Two Baselines

In Table 2 of our main paper, we present the overall performance of the proposed training-free and fine-tuning baselines on the test set. Here, we further report their performance on each individual event class. As shown in Fig. A1, the fine-tuning baseline outperforms the training-free baseline in most event classes (approximately 56 out of 67) across all evaluation metrics. This highlights the benefits of additional fine-tuning on training data. Moreover, we observe that some event classes, such as *slot machine* and *chicken crowing*, remain challenging for prediction, suggesting avenues for further improvement in future work.

G. More Details on Prompts for adapting Video-LLaMA2 to our OV-AVEL task

In Table 9 of our main paper, we compare the training-free baseline with an advanced audio-visual LLM, namely Video-LLaMA2 [2]. Video-LLaMA2 can process video frames and, more importantly, it can handle *general* audio signals that are not limited to human speech, unlike other audio-visual LLMs [5]. This makes it particularly suitable for the studied OV-AVEL task. Here, we provide more details on the prompts for adapting Video-LLaMA2 for the OV-AVEL task. Specifically, we tried several prompts and found the following prompt to be the most robust and effective for making predictions: "*Instruction: For the given 10-second video, divide it into 10 one-second segments. For each segment, if its audio and visual streams describe the same event, assign the label "x" as "1"; otherwise, label this segment as "0". User request: After processing all 10 video segments, you will obtain a list with 10 elements, each element being either "1" or "0" according to the above*

Instruction. Finally, return the most relevant event category of the video from the candidate category list: ["airplane flyby", "ambulance siren", "arc welding", "baby laughter", "basketball bounce", "bird chirping", "bowling impact", "cat purring", "cattle mooing", "chainsawing trees", "chicken crowing", ...(notebly, all event category texts should be listed; here, we omit the remaining ones for simplicity)]. The output format should be: "ave:" A python list [x, x, x, x, x, x, x, x, x, x] (replace "x" with "1" or "0" according to the prediction); Insert a line break. "class:" the most highly relevant class from the given category list (no punctuation needed at the end)." Readers may directly test this prompt on the official demo website using Hugging Face platform provided by authors of Video-LLaMA2 [2]: `https://huggingface.co/spaces/lixin4ever/VideoLLaMA2`. In this way, we can obtain the audio-visual event predictions of each test video and compare its performance with the proposed training-free baseline, as reported in Table 9 of our main paper. Additionally, we display some qualitative results in Fig. A2 and Fig. A3 and provide more discussions in Sec. H.

H. Qualitative Results

We finally present some intuitive video examples for OV-AVEL, as shown in Fig. A2 and Fig. A3. Specifically, we visualize the predictions generated by Video-LLaMA2 [2], along with the proposed training-free and fine-tuning baselines. As shown in the figures, the proposed fine-tuning baseline generally yields more accurate temporal localization results for both seen and unseen events/videos. For instance, in the three examples shown in Fig. A2, Video-LLaMA2 tends to predict most video segments as *background*, indicating its limitation in accurately perceiving the audio-visual correspondence at a fine-grained temporal-level. Although the training-free baseline performs better than Video-LLaMA2, the predictions for some video segments remain unsatisfactory. In contrast, the fine-tuning baseline performs better in localizing temporal segments containing audio-visual events and classifying the event categories. Similar phenomena can also be observed from Fig. A3. These qualitative results, along with the quantitative results presented in our main paper, suggest the effectiveness and superiority of the proposed fine-tuning baseline.

References

[1] Peijun Bao, Wenhan Yang, Boon Poh Ng, Meng Hwa Er, and Alex C Kot. Cross-modal label contrastive learning for unsupervised audio-visual event localization. In *AAAI*, pages 215–222, 2023. 2

[2] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 3

[3] Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. Learning event-specific localization preferences for audio-visual event localization. In *ACM MM*, pages 3446–3454, 2023. 2

[4] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 2

[5] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[7] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 2

[8] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pages 1–5, 2023. 2

[9] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *CVPR*, pages 19989–19998, 2022. 2

[10] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. MM-Pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *ACM MM*, pages 6241–6249, 2022.

[11] Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *TPAMI*, pages 1–18, 2022. 2

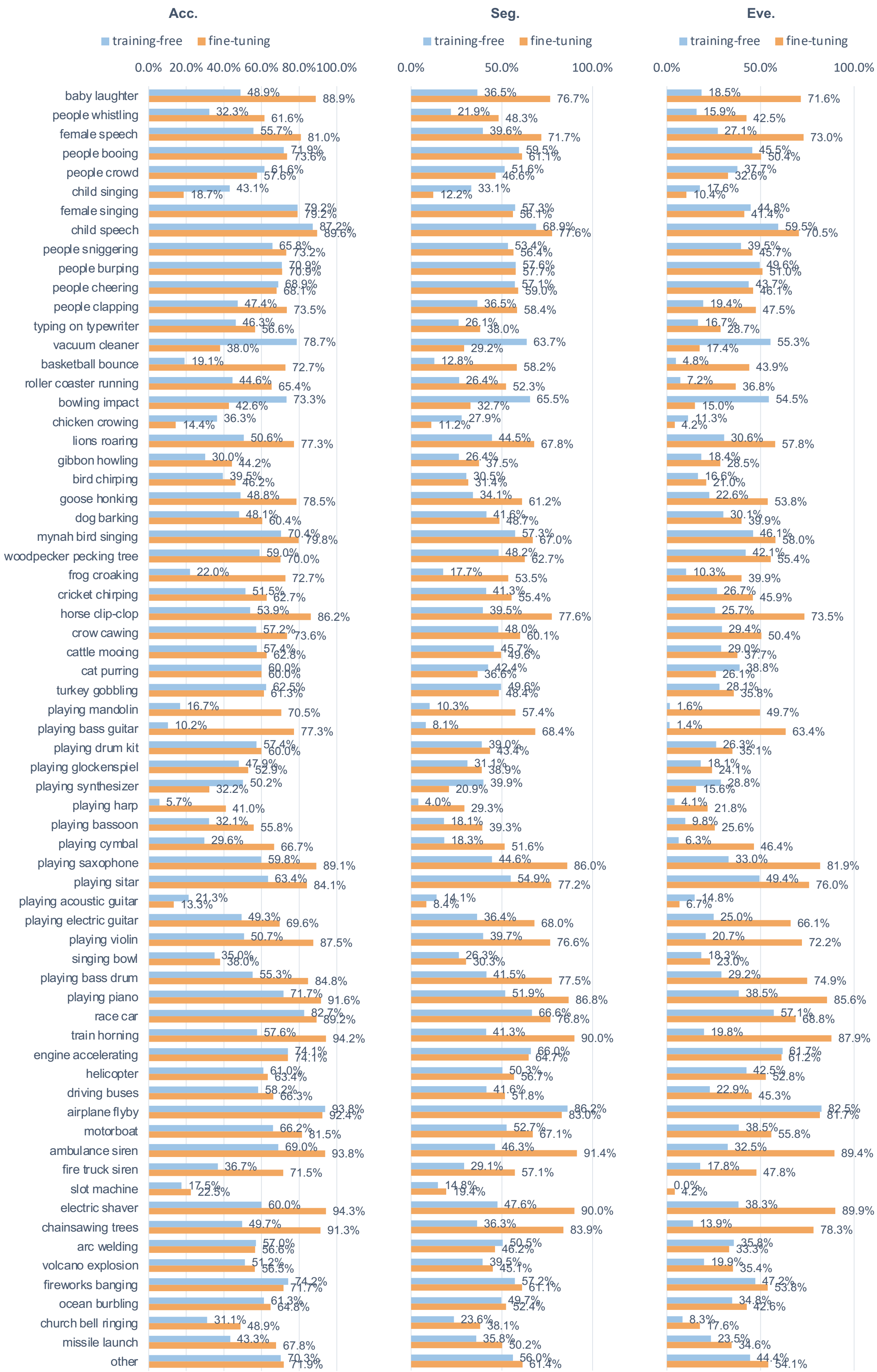

Figure A1. **Detailed performance of the proposed two baselines on each event class.**

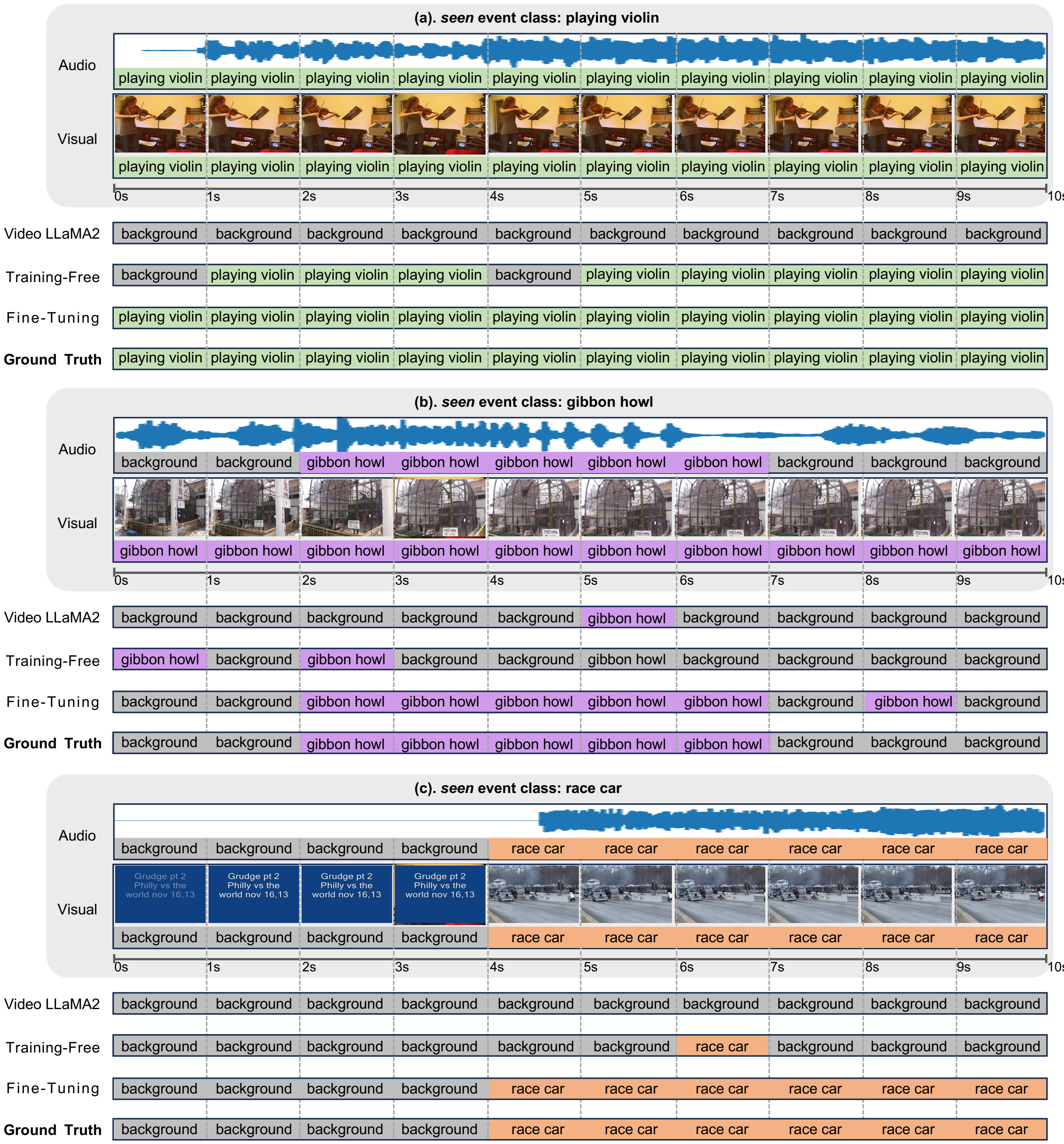

Figure A2. **Qualitative examples for seen audio-visual event localization.**

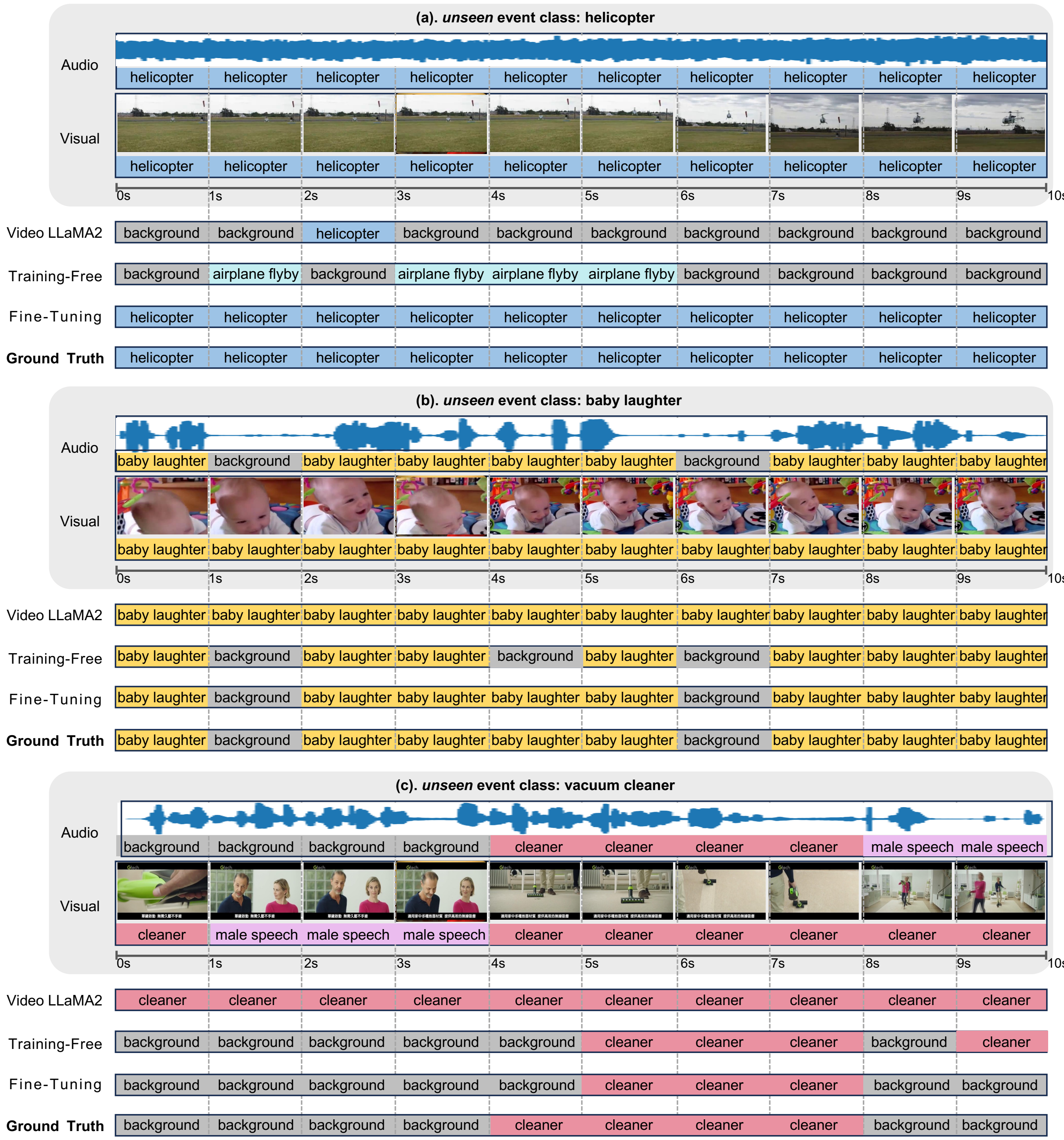

Figure A3. **Qualitative examples for unseen audio-visual event localization.**