Supplementary Material: Towards Omnimodal Expressions and Reasoning in Referring Audio-Visual Segmentation

Kaining Ying Henghui Ding✉ Guangquan Jie Yu-Gang Jiang
Fudan University, China
https://henghuiding.com/OmniAVS/

A. Benchmark: OmniAVS

A.1. Category Distribution

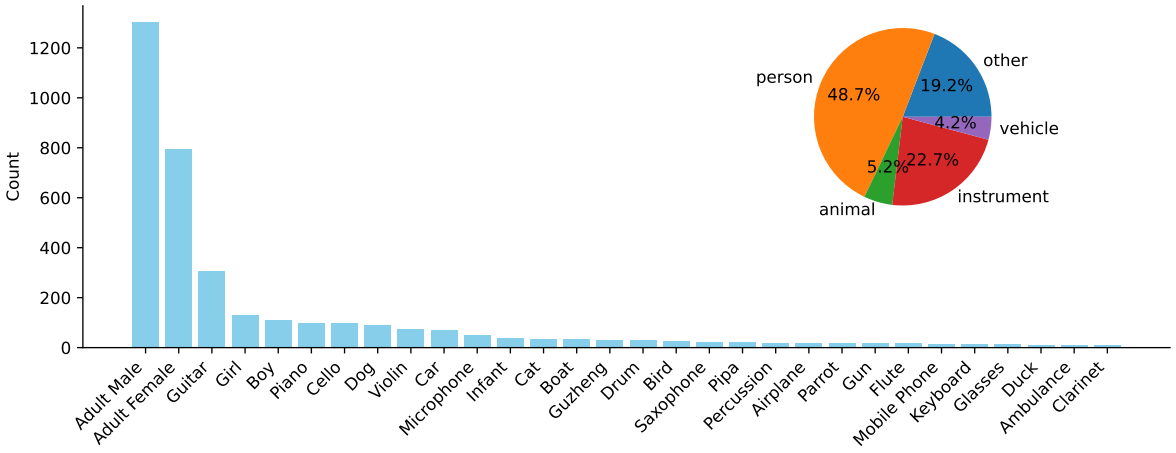

Figure I. Category distribution of OmniAVS.

As shown in Figure I, with the person category accounting for 48.7% of all instances. The person category can be further broken down into adult males, adult females, boys, girls, children, elderly men, elderly women, and infants. The remaining categories include various musical instruments (guitar, piano, drums, *etc*.), electronic devices (phones, speakers), and other common objects (cars, animals). This diverse category distribution enables comprehensive evaluation of audio-visual segmentation capabilities across a wide range of real-world scenarios. The dataset captures rich interactions between objects and their corresponding audio signals, making it particularly suitable for audio-visual segmentation tasks.

A.2. Data Source

The videos in OmniAVS come from three main sources: **1)** 1,657 web videos collected from various online platforms, which cover diverse real-world scenes including concerts, sports events, street views, and nature scenes, as shown in Figure 1. **2)** 397 videos from TVQA dataset [15], which contain rich dialogues and human interactions in TV shows and movies, as demonstrated in Figure 6(a). **3)** 50 self-recorded videos capturing daily life scenarios such as cooking, playing instruments, and casual conversations. This diverse collection of video sources ensures our dataset covers a wide range of audio-visual scenarios, from professionally produced content to natural daily interactions, making it comprehensive for evaluating audio-visual segmentation models.

✉ Henghui Ding (henghui.ding@gmail.com) is the corresponding author with the Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China.

A.3. Word Analysis

Figure II. Word cloud of the top 100 words in the OmniAVS.

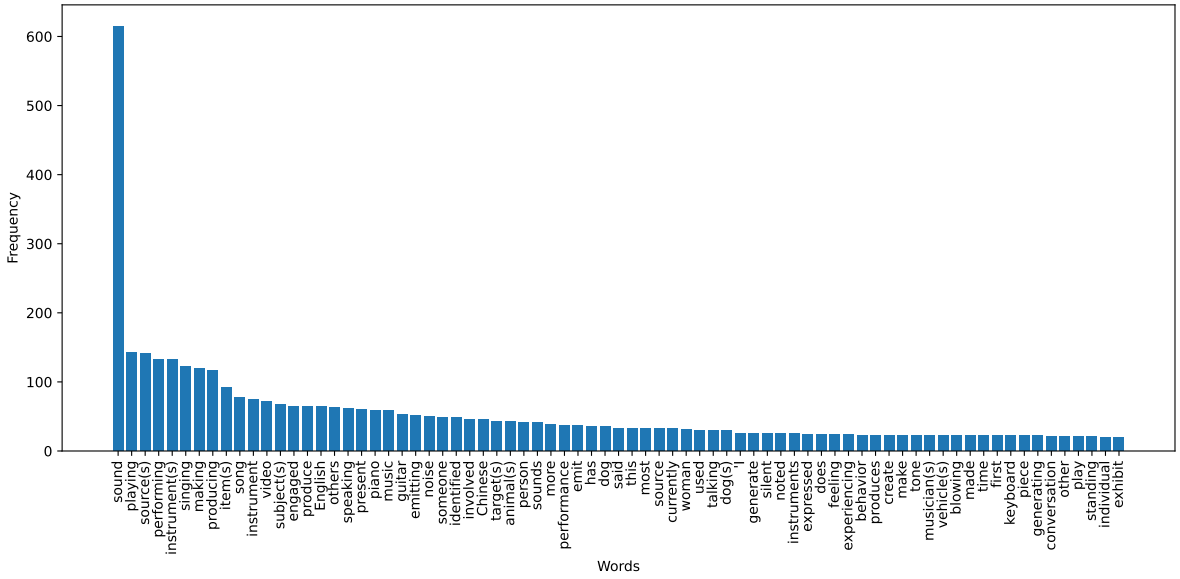

Figure III. Word frequency of the top 75 words in the OmniAVS.

We analyze the word distribution in our dataset through word cloud (Figure II) and word frequency statistics (Figure III). The most common words in the dataset are “sound”, “source”, “playing”, “performing”, *etc*., indicating that sound is a key focus of the dataset. Additionally, words describing audio such as “noise”, “English”, “Chinese” demonstrate the dataset’s attention to audio content. This word distribution analysis demonstrates that our dataset contains rich and diverse referring expressions that leverage multiple types of audiovisual attributes and relationships for object description and reference.

A.4. Annotation GUI

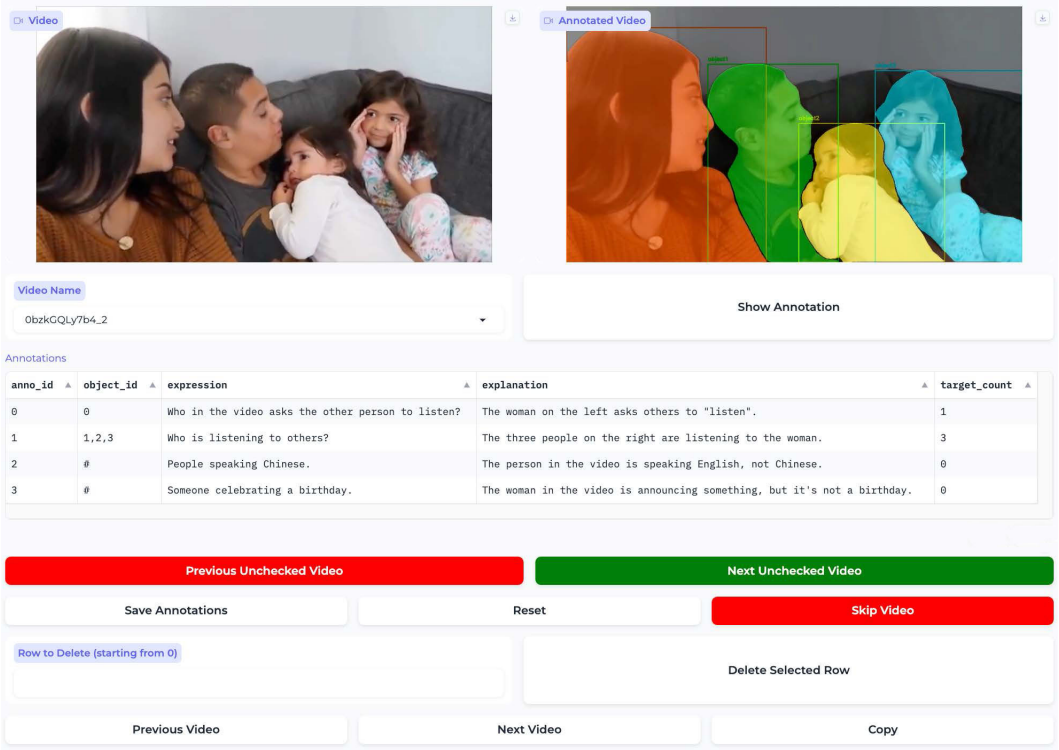

Figure IV. Screenshot of our expression annotation interface.

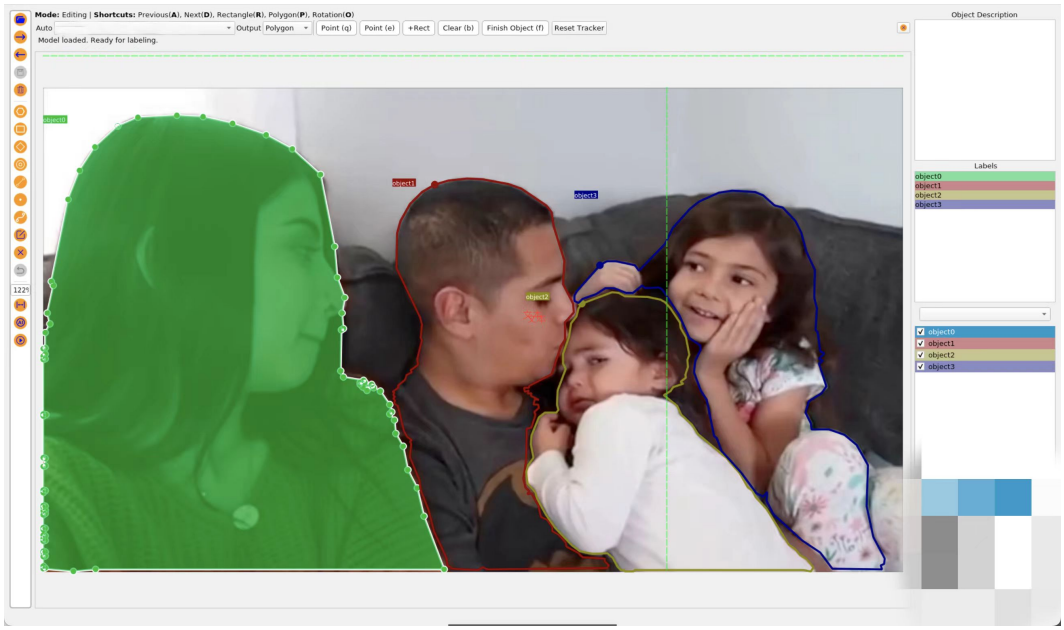

Figure V. Screenshot of our mask annotation interface.

In this appendix, we introduce the annotation platform used during dataset labeling. We use the interface shown in Figure IV to annotate expressions and the interface in Figure V to annotate corresponding masks. The mask annotation platform is inspired by [18].

A.5. Compare with More Datasets

Table I. Statistical comparison between the newly proposed OmniAVS and other datasets of related tasks. Avail., Expl., and Expr. are abbreviations for Availability of reasoning, Explanations, and Expressions, respectively.

Dataset	Venue	Content		Referring			Reasoning		Statistics					
		Video	Audio	Text	Audio	Image	Avail.	Expl.	Video	Frame	Object	Mask	Expr.	Expl.
J-HMDB Sentences [10]	[CVPR'18]	✓	✗	✓	✗	✗	✗	✗	928	928	928	-	928	✗
A2D Sentences [10]	[CVPR'18]	✓	✗	✓	✗	✗	✗	✗	3,782	11,936	4,825	-	6,656	✗
Refer-DAVIS-2016 [12]	[ACCV'18]	✓	✗	✓	✗	✗	✗	✗	50	3,455	50	-	100	✗
Refer-DAVIS-2017 [12]	[ACCV'18]	✓	✗	✓	✗	✗	✗	✗	90	13,543	205	-	1,544	✗
Refer-YouTube-VOS [16]	[ECCV'20]	✓	✗	✓	✗	✗	✗	✗	3,975	116,523	7,451	131k	15,009	✗
MeViS [6]	[ICCV'23]	✓	✗	✓	✗	✗	✗	✗	2,006	44,300	8,175	443k	28,570	✗
ReVOS [20]	[ECCV'24]	✓	✗	✓	✗	✗	✓	✗	1,042	116,321	5,535	469k	35,074	✗
ReasonSeg [14]	[CVPR'24]	✓	✗	✓	✗	✗	✓	✗	1,218	1,218	1,218	-	7,308	239
Flickr-SoundNet [2]	[CVPR'18]	✓	✓	✗	✗	✗	✗	✗	5,000	5,000	-	-	-	✗
VGG-SS [5]	[CVPR'21]	✓	✓	✗	✗	✗	✗	✗	5,158	5,158	-	-	-	✗
AVSBench [22]	[ECCV'22]	✓	✓	✗	✗	✗	✗	✗	12,356	82,972	13,500	-	-	✗
Ref-AVS [19]	[ECCV'24]	✓	✓	✓	✗	✗	✗	✗	4,002	40,020	6,888	78k	20,261	✗
OmniAVS (ours)	[ICCV'25]	✓	✓	✓	✓	✓	✓	✓	2,104	103,087	4,277	206k	61,095	34,841

As shown in Table I, we provide a comprehensive comparison with more datasets [2, 5, 6, 10, 12, 14, 16, 19–22] from related tasks. Early datasets like J-HMDB Sentences [10] and A2D Sentences [10] only provide text expressions for a small number of videos without mask annotations. Refer-DAVIS [12] series provide high-quality mask annotations but are limited in scale. Recent referring video segmentation datasets like Refer-YouTube-VOS [16], MeViS [6], and ReVOS [20] significantly expand the scale but still focus on silent videos with only text expressions [8]. ReasonSeg [14] introduces reasoning capability and explanations but lack of multimodal expressions. While many RVOS datasets inherit mask annotations from existing video object segmentation datasets [7, 9], we provide newly annotated masks. Audio-visual datasets like Flickr-SoundNet [2] and VGG-SS [5] only provide bounding box annotations on a small number of frames. AVSBench [22] expands the scale but focuses on sound-emitting objects without referring expressions. Ref-AVS [19] introduces text expressions to audio-visual segmentation but lacks reasoning capability and explanations. In contrast, our OmniAVS dataset uniquely combines multimodal content (audio-visual videos), diverse expression modalities (text, speech, sound, image), reasoning capability with explanations, and comprehensive mask annotations, making it a more complete benchmark for multimodal video understanding and segmentation.

A.6. More Examples in OmniAVS

We provide more examples from our dataset in Figures VI to VIII. For more details, please refer to `visualizations.mp4` in the supplementary materials.

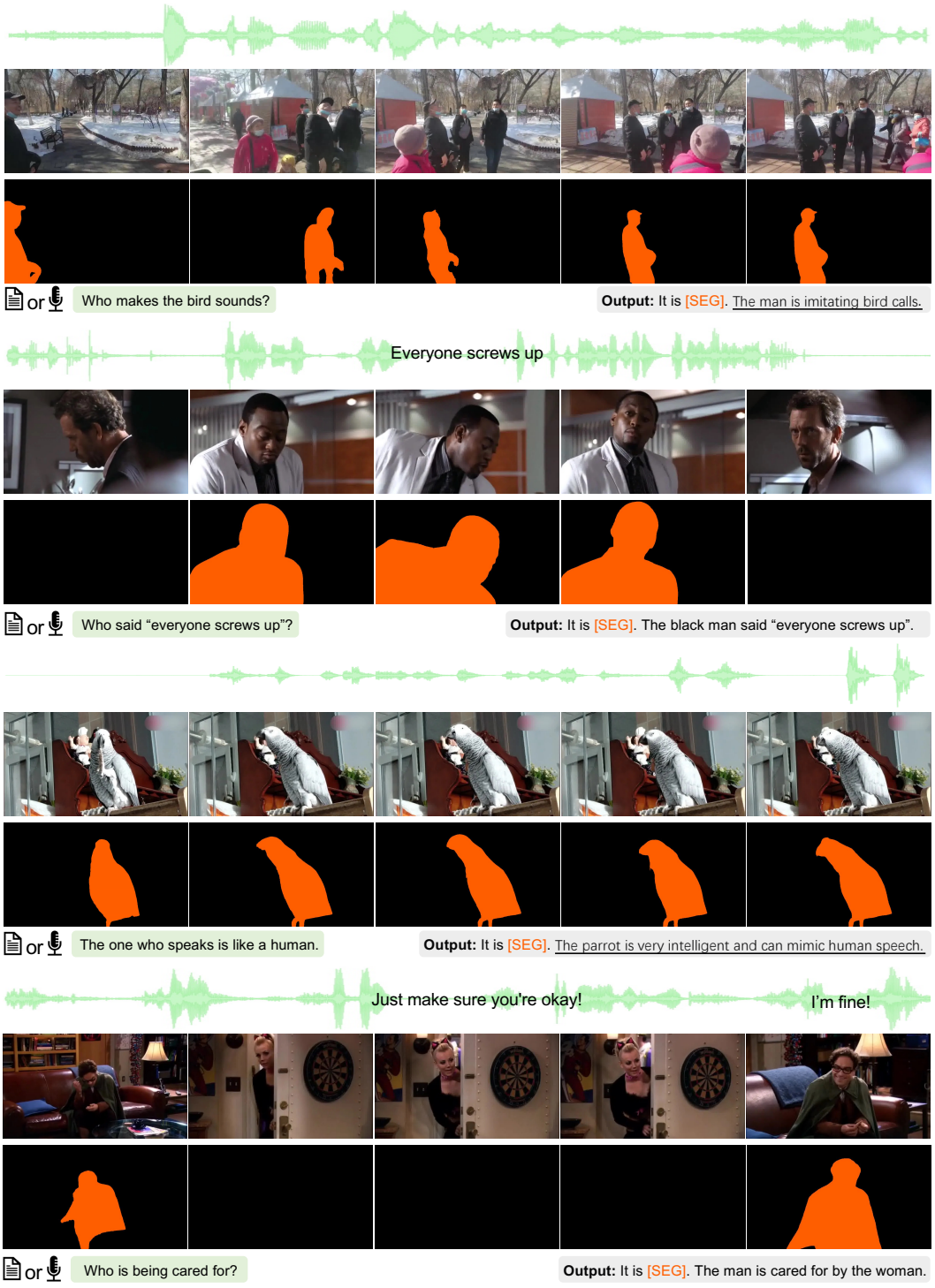

Figure VI. More examples from our OmniAVS dataset: text or speech instructions.

Figure VII. More examples from our OmniAVS dataset: text or speech instructions with sound.

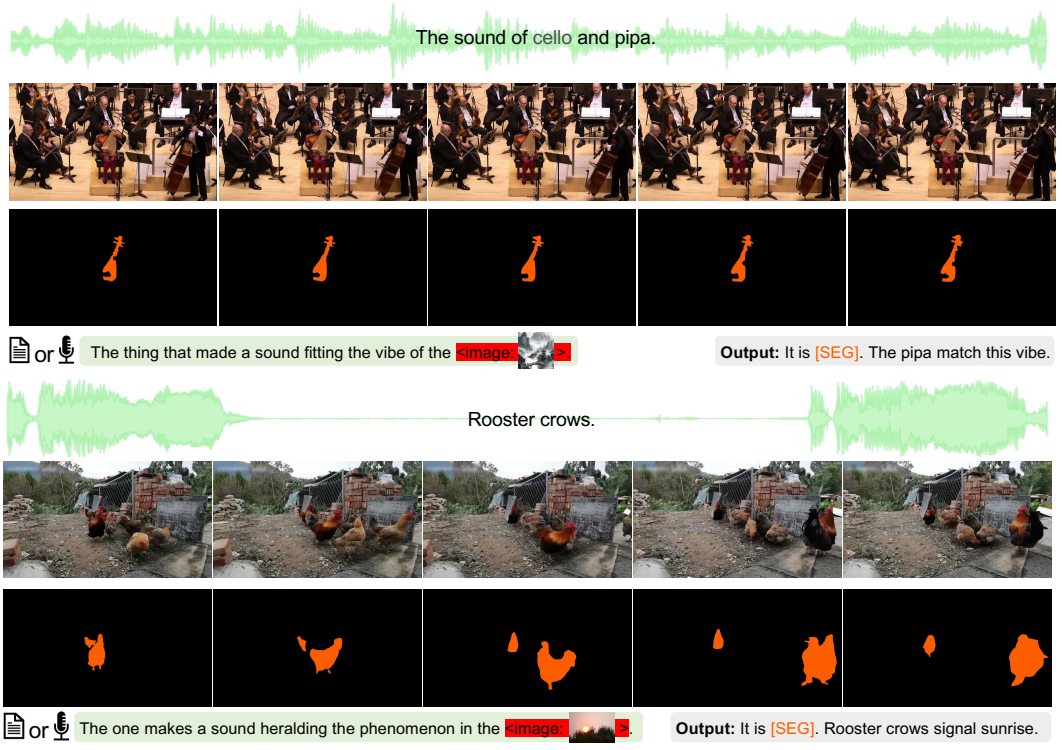

Figure VIII. More examples from our OmniAVS dataset: text or speech instructions with image.

B. Model: OISA

B.1. Learning Objectives

We employ multiple loss functions to train our model. The total loss $\mathcal{L}$ consists of three components:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}^{\text{text}} + \lambda_2 \mathcal{L}_{\text{DICE}}^{\text{mask}} + \lambda_3 \mathcal{L}_{\text{BCE}}^{\text{mask}} + \lambda_4 \mathcal{L}_{\text{BCE}}^{\text{label}}, \tag{i}$$

where $\mathcal{L}_{\text{CE}}^{\text{text}}$ is the cross entropy loss for text generation, $\mathcal{L}_{\text{DICE}}^{\text{mask}}$ is the DICE loss and $\mathcal{L}_{\text{BCE}}^{\text{mask}}$ is the binary cross entropy loss for mask prediction, and $\mathcal{L}_{\text{BCE}}^{\text{label}}$ is the binary cross entropy loss for mask existence classification. We set $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\lambda_3 = 2$, and $\lambda_4 = 1$ as the corresponding loss weights.

B.2. Training Data

Table II. Training data statistics for audio-text alignment.

Task	Dataset	Samples
Automatic Speech Recognition	GigaSpeech [4]	301,723
	CommonVoice [3]	4,063
Automatic Audio Caption	Auto-ACD [17]	1,855,829
	AudioCaps [13]	38,889
	MusicCaps [1, 11]	4,753

As shown in Table II, we use two types of datasets for audio-text alignment training:

1. **Automatic Speech Recognition (ASR) datasets**: We use GigaSpeech [4] and CommonVoice [3], which contain speech-to-text pairs. For ASR training, we use different prompts, such as (1) *"Please convert this speech into text."*, (2) *"What is said in this speech clip?"*, (3) *"Please convert this speech into text."*, and (4) *"What is the textual representation of this speech?"*. GigaSpeech provides 301,723 samples while CommonVoice contributes 4,063 samples.
2. **Automatic Audio Caption (AAC) datasets**: We leverage Auto-ACD [17], AudioCaps [13], and MusicCaps [1] built upon AudioSet [11]. Auto-ACD is the largest with 1.86M samples, followed by AudioCaps with 39K samples and MusicCaps with 4.8K samples. For AAC training, we use different prompts, such as (1) *"Please describe what you hear in this audio."*, (2) *"What is happening in this audio clip?"*, (3) *"Describe the audio."*, and (4) *"What can you tell me about this audio?"*. These datasets contain audio clips paired with descriptive captions.

In total, we use around 2.2M audio-text pairs for training the audio-text alignment module.

We introduced the dataset for the omnimodal instructed segmentation tuning phase in Section 5.1, where we use a unified prompt template *"Please segment the object this sentence describes: expression"*. If there is a target object exists, the answer is *"Sure, it is [SEG]."*, and if there is no target, the response is *"No target matches this expression."*

References

[1] Andrea Agostinelli, Timo Lee, Emanuele Bugliarello, Lukas Mitterhofer, James Whitehead, Jaesung Choi, and Jesse Engel. MusicCaps: A Music Audio Captioning Dataset. https://huggingface.co/datasets/google/MusicCaps, 2023. 7

[2] Relja Arandjelovic and Andrew Zisserman. Look, Listen and Learn. In *Int. Conf. Comput. Vis.*, 2017. 4

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. In *Lang. Resour. Eval. Conf.*, 2020. 7

[4] Guoguo Chen, Shuzhou Chai, et al. GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*, 2021. 7

[5] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing Visual Sounds the Hard Way. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 4

[6] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *Int. Conf. Comput. Vis.*, 2023. 4

[7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A New Dataset for Video Object Segmentation in Complex Scenes. In *Int. Conf. Comput. Vis.*, 2023. 4

[8] Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *arXiv*, 2025. 4

[9] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Yu-Gang Jiang, Philip HS Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv*, 2025. 4

[10] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and Action Video Segmentation from a Sentence. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 4

[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017. 7

[12] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video Object Segmentation with Language Referring Expressions. In *ACCV*, 2019. 4

[13] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. In *Conf. North Am. Chapter Assoc. Comput. Linguist.*, 2019. 7

[14] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 4

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, Compositional Video Question Answering. In *Proc. of the Conf. on Empirical Methods in Nat. Lang. Process.*, 2018. 1

[16] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In *Eur. Conf. Comput. Vis.*, 2020. 4

[17] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. Auto-ACD: A Large-scale Dataset for Audio-Language Representation Learning. In *ACM Int. Conf. Multimedia*, 2024. 7

[18] Wei Wang. Advanced Auto Labeling Solution with Added Features. https://github.com/CVHub520/X-AnyLabeling, 2023. 3

[19] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes. In *Eur. Conf. Comput. Vis.*, 2024. 4

[20] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning Video Object Segmentation via Large Language Models. In *Eur. Conf. Comput. Vis.*, 2024. 4

[21] Kaining Ying, Hengrui Hu, and Henghui Ding. MOVE: Motion-guided few-shot video object segmentation. In *Int. Conf. Comput. Vis.*, 2025.

[22] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-Visual Segmentation. In *Eur. Conf. Comput. Vis.*, 2022. 4