

Supplementary Materials: Open-Vocabulary Audio-Visual Semantic Segmentation

Anonymous Authors

1 DATASET STATISTIC

To facilitate a comprehensive evaluation of OV-AVSS, we partition a novel dataset called **AVSBench-OV**, derived from AVSBench-Semantic [2]. AVSBench-Semantic encompasses 70 categories, incorporating both single and multiple sound source scenarios. Each video within AVSBench-Semantic is truncated to either 5 or 10 seconds in duration, with one frame extracted per second. Following the partitions in LVIS [1], we split the categories within AVSBench-OV into 40 base categories representing those seen during training and inherited from frequent and common categories, and 30 novel (unseen and unheard) categories disjoint from the base categories. To ensure consistency, we eliminated sample videos from the training subset whenever a novel category appeared in annotations, thereby restricting the training data exclusively to base categories.

As illustrated in Fig. 1, we show the category names in the training set along with the video numbers for each category. We reserve 5,184 videos with 40,095 frames from the AVSBench-Semantic dataset and videos comprising across 40 base categories. We also display the statistics of the testing set in Fig. 2. There are 1,490 videos consisting of 40 base categories and 30 novel categories.

2 TEXT PROMPTS

Since the open-vocabulary classification model is trained on full sentences, we feed the category names into a prompt template first, and use an ensemble of various prompts. Our list of prompt templates is shown below:

- “a photo of a .”
- “This is a photo of a ”
- “There is a in the scene”
- “There is the in the scene”
- “a photo of a in the scene”
- “a photo of a small ”
- “a photo of a medium .”
- “a photo of a large .”
- “This is a photo of a small .”
- “This is a photo of a medium .”
- “This is a photo of a large .”
- “There is a small in the scene.”
- “There is a medium in the scene.”
- “There is a large in the scene.”

REFERENCES

- [1] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5356–5364.
- [2] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. 2023. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190* (2023).

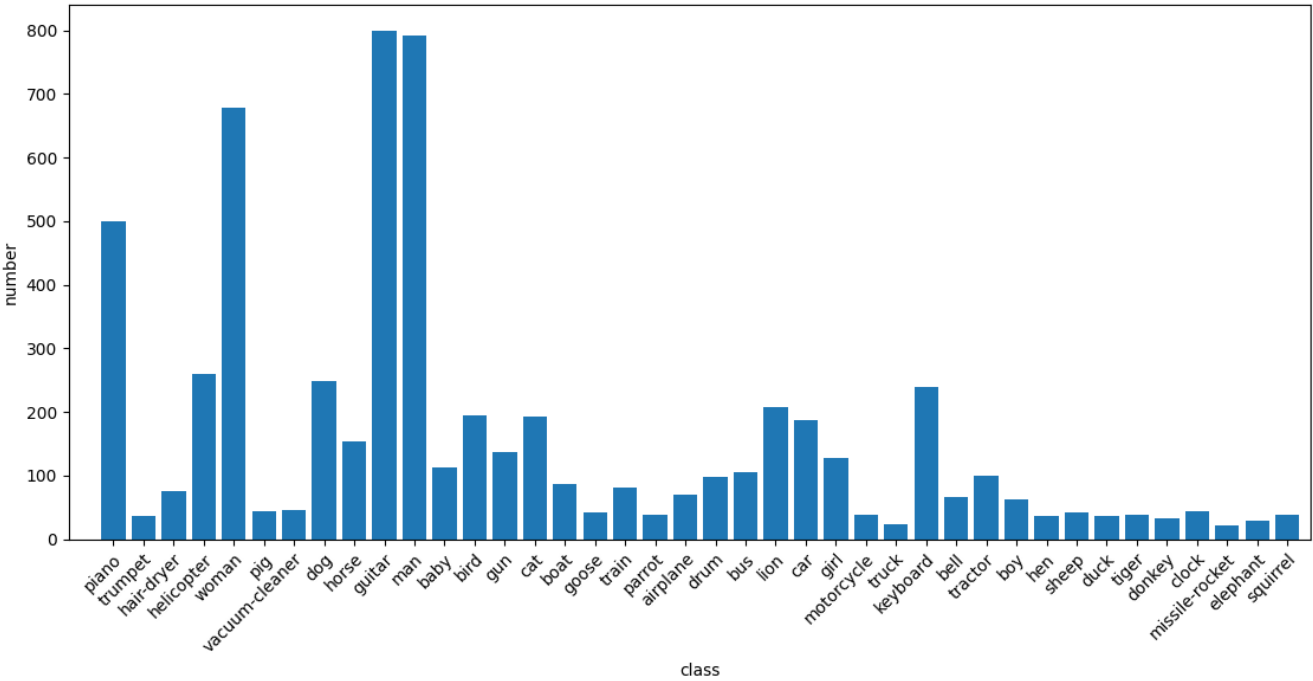


Figure 1: Statistics of the training set of AVSBench-OV, which consists of 40 base categories.

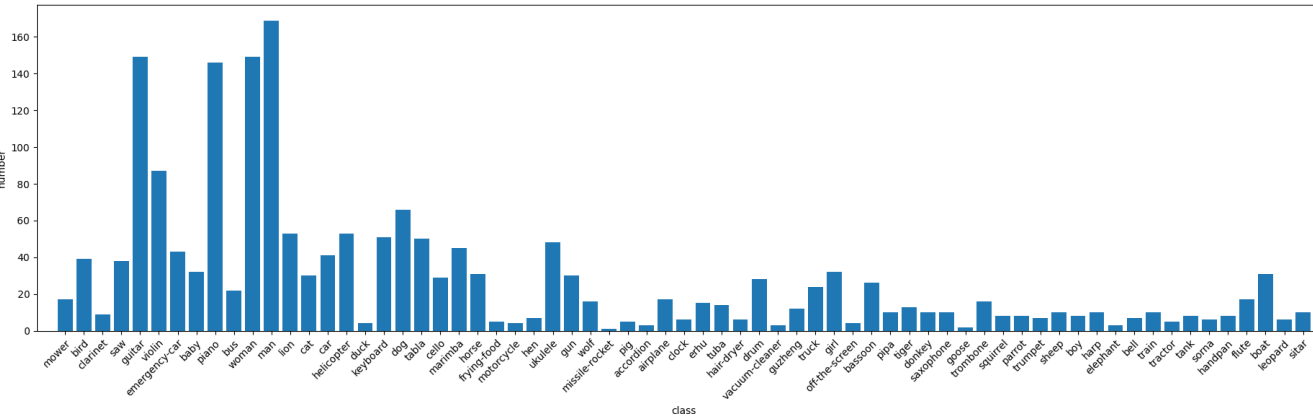


Figure 2: Statistics of the testing set of AVSBench-OV, which consists of 40 base categories and 30 novel categories.