QDFormer: Towards Robust Audiovisual Segmentation in Complex Environments with Quantization-based Semantic Decomposition

Xiang Li[1]*, Jinglu Wang[2], Xiaohao Xu[3], Xiulian Peng[2], Rita Singh[1], Yan Lu[2], Bhiksha Raj[1]
[1] CMU, [2] Microsoft Research Asia, [3] UMich

A. More Experiments

Pvt-v2 [5]	Object-M	Sementic
	$\mathcal{J}\&\mathcal{F}$	mIoU
AVSBench [6]	59.3	29.8
AVSegFormer [1]	63.8	42.0
CATR [3]	64.5	32.8
Ours	65.8	54.5

Table A. Performance with larger backbone.

Larger backbone. We demonstrate the performance on AVSBench test set with Pvt-v2 [5] backbone as shown in Tab. A. We notice that our method consistently outperforms previous baselines CATR [3] and AVSegFormer [1] on both AVS-Object-Multi and AVS-Semantic datasets.

frame number	Object-M	Sementic
	$\mathcal{J}\&\mathcal{F}$	mIoU
3	60.9	45.4
5	61.6	46.6
7	-	46.6

Table B. Ablation on the input frame number.

Frame number. We ablate the influence of input frame number during training. As shown in Tab. B, we notice a frame number of five achieves the best performance. For the AVS-Object dataset, since the maximum clip length is five, we do not experiment with larger frame number. Please note that the frame number is only fixed during training and the model can accept arbitrary frame numbers during inference.

Transformer decoder layer number. We conduct an ablation study on transformer decoder layer numbers in semantic decoders. As shown in Tab. C, a transformer decoder layer of 3 achieves the best performance. We notice that even a single-layer transformer decoder for semantic decomposition can lead to a good performance.

*This work was done when Xiang Li and Xiaohao Xu were interns at Microsoft.

layer number	Object-M	Sementic
	$\mathcal{J}\&\mathcal{F}$	mIoU
1	61.3	45.6
3	61.6	46.6
5	61.0	46.0

Table C. Ablation on transformer decoder layer number.

frame resolution	Sementic
	mIoU
224×	46.6
640×	49.2

Table D. Ablation on input frame resolution.

Input Resolution. The default setting of AVSBench is 224×224 (following the sound source localization convention) for both AVS-Object and AVS-Semantic datasets. While AVS-Semantic actually provides high-resolution (720p) frames. We conduct experiments to ablate the input resolution to facilitate future comparison. Following the semantic segmentation convention, we scale the input frames to the longest side 224 or 640. The results are illustrated in Tab. D. We only conduct ablation on AVS-Semantic since the resolution of AVS-Object is low-resolution (224×224). The results are reported with the ResNet-50 backbone.

Token Number	Object-M $\mathcal{J}\&\mathcal{F}\uparrow$	Sementic mIoU↑
1	59.7	40.2
3	61.0	43.5
5	61.6	46.6
7	61.6	45.9
9	61.2	46.3

Table E. Ablation on decomposed token number.

Ablation on semantic token number. We ablate the semantic token number for the global-ASD and local-ASD in Tab. E. We observe that a token number of 5 yielded the best performance. This can be attributed to the fact that the max-

imum number of mixed sound sources for the audio-visual dataset is 5.

Codebook Size	Object-M $\mathcal{J}\&\mathcal{F}\uparrow$	Sementic mIoU↑
1	52.7	24.8
32	61.4	31.5
64	60.0	43.2
128	61.6	46.6
256	60.6	46.1

Table F. Ablation on codebook size.

Ablation on codebook size. The cardinality of the codebook is essential to our semantic decomposition. Ideally, we aim to constrain the cardinality of the codebook to be close to the semantic category number. We ablate on codebook size from 1 to 256. When the codebook size equals 1, all the decomposed audio tokens are the same, resulting in all the same segmentation results. As shown in Tab. F, we notice even a codebook size of 1 achieves 24.8 mIoU on AVS-Semantic. A codebook of size=128 achieves the best performance. Please note that a codebook size slightly larger than the category number, e.g.128, will not hamper the semantic decomposition capability of our method, since $128 \ll 70^N$ where $N > 1$ is the maximum sound source number, and 70 is the category number.

Importance of the single-source audio on the semantic decomposition of multi-source audio representation.

We present empirical evidence that the single-source audio samples significantly contribute to the success of semantic decomposition. To demonstrate this, we compare the performance of our model trained on two training sets with the same number of samples: one contains solely multi-source audio samples, and the other contains single- and multi-source audio samples with a ratio of 1:1. As illustrated in Fig. B, the model trained solely on multi-source audio samples exhibits inferior performance compared to the model trained on both types of samples, regardless of the token number and codebook size. We conjecture that the single-source samples serve as informative anchors that assist the model in learning the correct distributions of the decomposed simplex spaces for multi-source samples. In the absence of single-source samples, the decomposition task could be more difficult due to the absence of such informative anchors.

Per-class IOU analysis. As is shown in Fig. A, we show the per-class iou score on the AVS-Semantic dataset. Our model demonstrates strong audio-guided segmentation capabilities for common head classes such as 'background', 'train', 'airplane', 'hair-dryer' and 'clock'. These classes are accurately segmented with a high level of precision and reliability. The model effectively distinguishes the 'background' class, providing a solid foundation for identifying and isolating foreground objects. It accurately segments transportation-related classes like 'train', 'airplane', and 'bus' capturing their intricate details and boundaries. Similarly, it excels in segmenting objects such as 'hair-dryer', 'clock' and 'tabla,' effectively separating them from the background. Even for more complex and nuanced classes like 'wolf,' our model demonstrates commendable segmentation performance, accurately delineating the contours and shape of the subject. Overall, our model showcases its ability to segment these common head classes with high accuracy and proficiency, making it a reliable choice for various segmentation tasks.

However, the scarcity of data samples for tail classes like 'utv', 'parrot', 'missile-rocket', 'harmonica', 'clipper', 'boy' and 'ax' in the presence of a long tail distribution can significantly impact the performance of our model, specifically in the task of segmentation. With limited examples to learn from, the model finds it challenging to capture the intricate patterns and unique characteristics associated with these classes. Consequently, the accuracy and reliability of segmentation results for the tail classes may be compromised, leading to suboptimal performance in accurately identifying and delineating these objects or entities of interest.

B. More Visualization & Video Demo

More qualitative results on AVS-Object. In our study, we provide visualizations of the qualitative results on AVS-Object, as shown in Fig. C. We compare our method with the approach proposed by Zhou et al. [7] and observe a notable difference in performance. Specifically, in the third frame of the video clip, the method proposed by Zhou et al. suffers from the false-positive problem, incorrectly segmenting objects. In contrast, our method consistently and accurately segments the correct objects throughout the entire video clip, demonstrating superior performance. Additionally, our method showcases better mask quality, with more precise and detailed segmentation boundaries. These results highlight the effectiveness and robustness of our approach in achieving accurate object segmentation in audio-visual scenes.

More qualitative results on AVS-Semantics. As is shown in Fig. D, Fig. E, Fig. F and Fig. G, our model exhibits exceptional proficiency in accurately segmenting both multiple and tiny sounding objects, showcasing its versatility and robustness in audio-guided segmentation tasks. Through the implementation of a decomposed and discretized audio representation, our model effectively captures the distinct acoustic characteristics of various objects, enabling precise delineation of multiple simultaneous sound sources. Furthermore, the model demonstrates remarkable capability in capturing the intricate details and nuances of tiny sounding

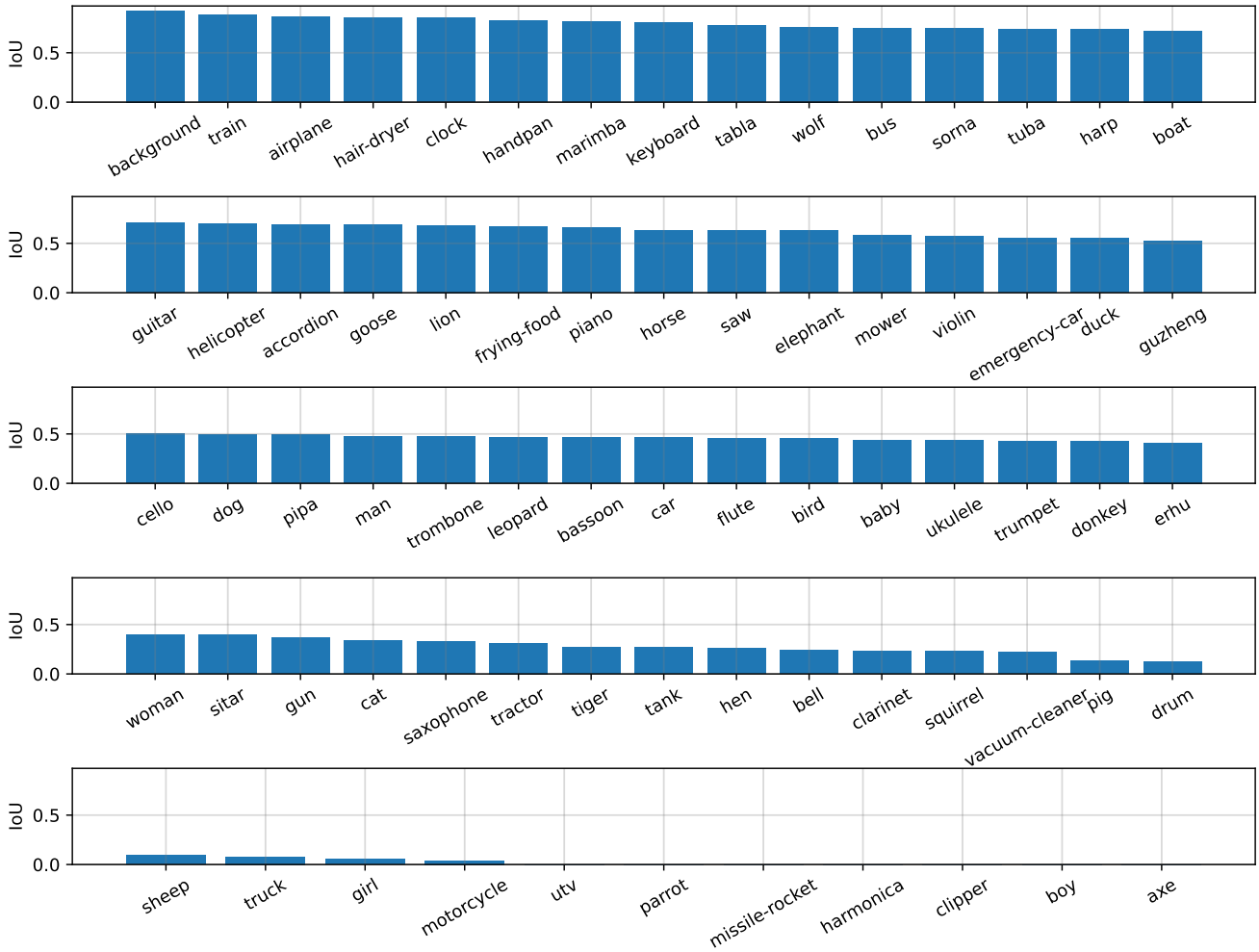

Figure A. **Per-class IOU Analysis.** Our model demonstrates strong audio-guided segmentation capabilities for common head classes, accurately capturing 'background', 'train', 'airplane', 'hair-dryer', and 'clock' with high precision. However, the limited data samples for tail classes like 'utv', 'parrot', 'missile-rocket', 'harmonica', 'clipper', 'boy', and 'ax' due to a long tail distribution adversely affect the model's segmentation performance, hindering accurate identification and delineation of these classes.

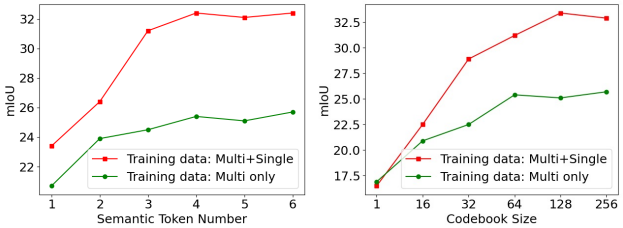

Figure B. Comparison of training w. and w./o. single-source data.

objects, ensuring accurate segmentation outcomes even for the smallest entities.

Video demo (with audio). We strongly recommend viewing the demo video provided in the supplementary materials, ensuring that you enable audio playback. Watching the video with audio will provide a comprehensive understanding of our audio-visual segmentation application, showcasing how our model utilizes a decomposed and discretized representation to achieve precise audio-visual segmentation results.

C. More Implementation Details

We set the $\lambda_{cls} = 2$, $\lambda_{L1} = 5$, $\lambda_{giou} = 2$, $\lambda_{dice} = 2$, $\lambda_{focal} = 5$, $\lambda_{com} = 0.5$ and $\lambda_{quant} = 1$ during all training process. A mask confidence threshold of 0.5 and a class confidence threshold of 0.1 is leveraged to filter out low-confident predictions. $C_v = C_e = C_q = 256$ is utilized. The positional embedding added in the transformers is the standard triangle positional embedding used in [4]. We set the layer number to three for all the transformers decoders (including local ASD, global ASD and TrD_{segm} in mask

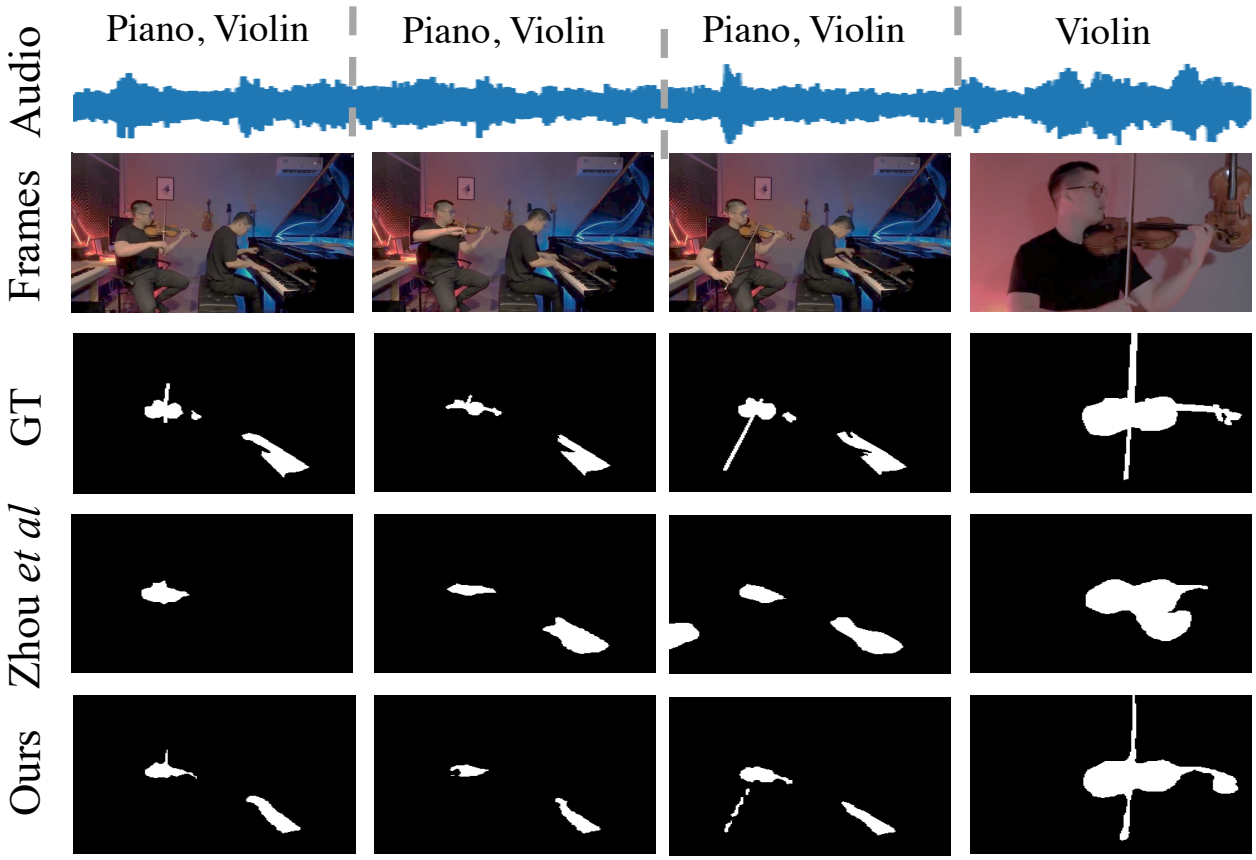

Figure C. Qualitative comparison to Zhou et al. [7] on AVS-Object. Our method outperforms Zhou et al.'s approach by consistently and accurately segmenting the correct objects throughout the entire video clip, showcasing superior performance and better mask quality. These results emphasize the effectiveness and robustness of our approach in achieving accurate object segmentation in audio-visual scenes.

decoder).

C.1. Encoders

Visual encoder. We extract frame-level visual features from each frame I_t with a shared backbone. The T extracted features are then fed into the deformable transformer encoder to further conduct temporal aggregation. Let us denote the extracted visual features as $F_v = \{f_t\}_{t=1}^{T}$, where $f_t \in \mathbb{R}^{C_v \times H \times W}$, and C_v, H, W denote the channel, height, width of the feature.

Acoustic encoder. We use VGGish [2] to extract audio features. Let the extracted audio feature be $F_a \in \mathbb{R}^{C_a \times L_a}$ where C_a is the dimension of acoustic feature space, and L_a is the audio clip length. Note that audio and video frames are already synchronized, thus the length of the audio clip is the same as the length of the video clip.

D. More details about inference

To tackle scenarios where queried content keeps changing, we perform per-frame inference. For each time t, we assign a class to the pixel at $[h, w]$ by

$$\arg \max_{C \in \{1, \cdots, K\}} \sum_{i=1}^{N} P_{i,t}[C] M_{i,t}[h, w], \tag{1}$$

where $P_{i,t}[C]$ is the probability of class C. Note that $\arg\max$ does not include the "empty" category ($\emptyset$) as AVS requires each output pixel to belong to one semantic category.

References

[1] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. *arXiv preprint arXiv:2307.01146*, 2023. 1

[2] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 4

[3] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun

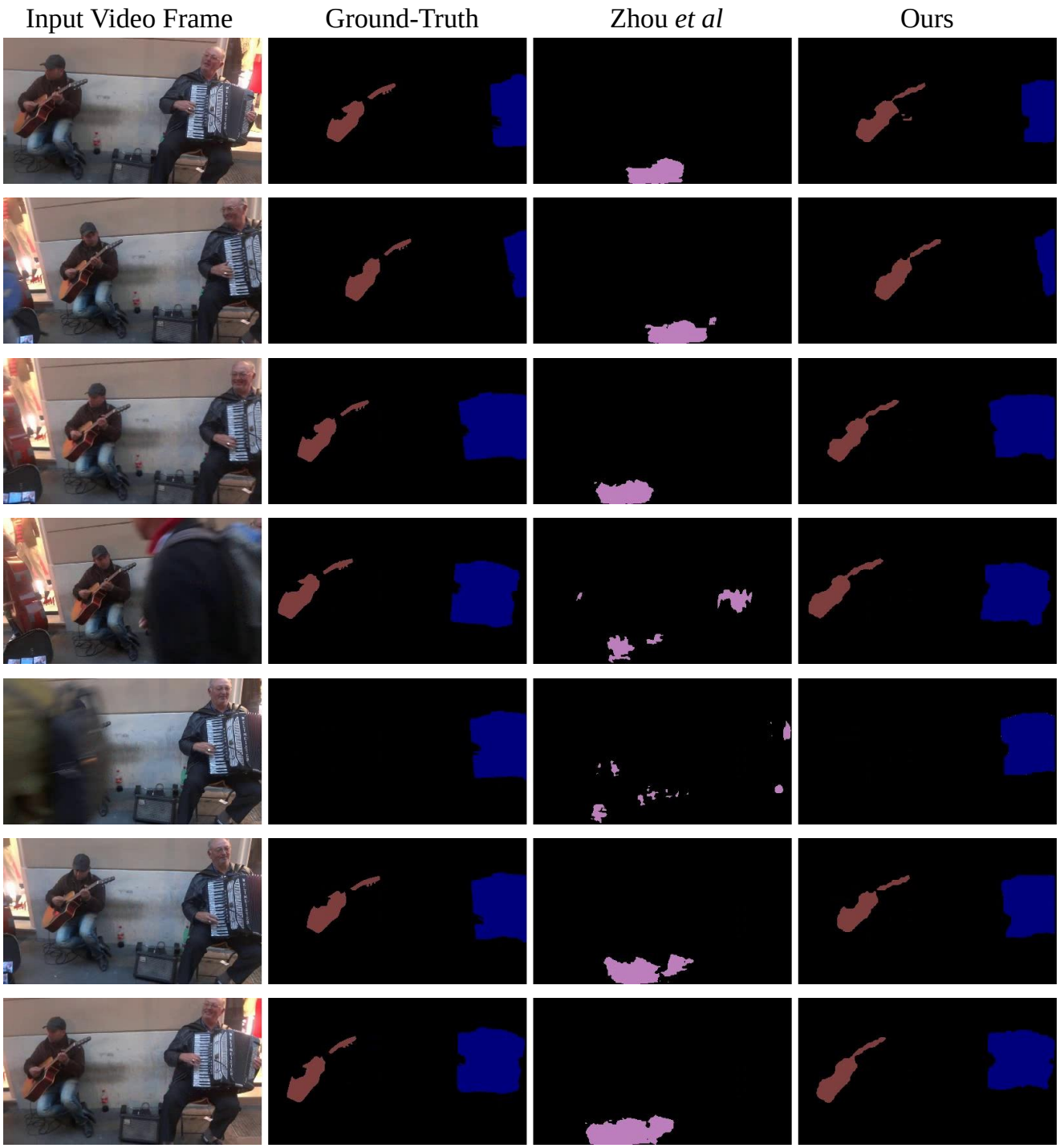

Figure D. Qualitative comparison to Zhou et al. [7] on AVS-Semantic. Each color represents a semantic category. Our model excels in accurately segmenting **multiple sounding objects**, showcasing its proficiency in audio-guided segmentation. This success can be attributed to the effective utilization of a decomposed and discretized audio representation, which enables the model to capture and analyze the distinct acoustic features of each object, resulting in precise segmentation outcomes.

Xun. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. *arXiv preprint arXiv:2309.09709*, 2023. 1

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkor-

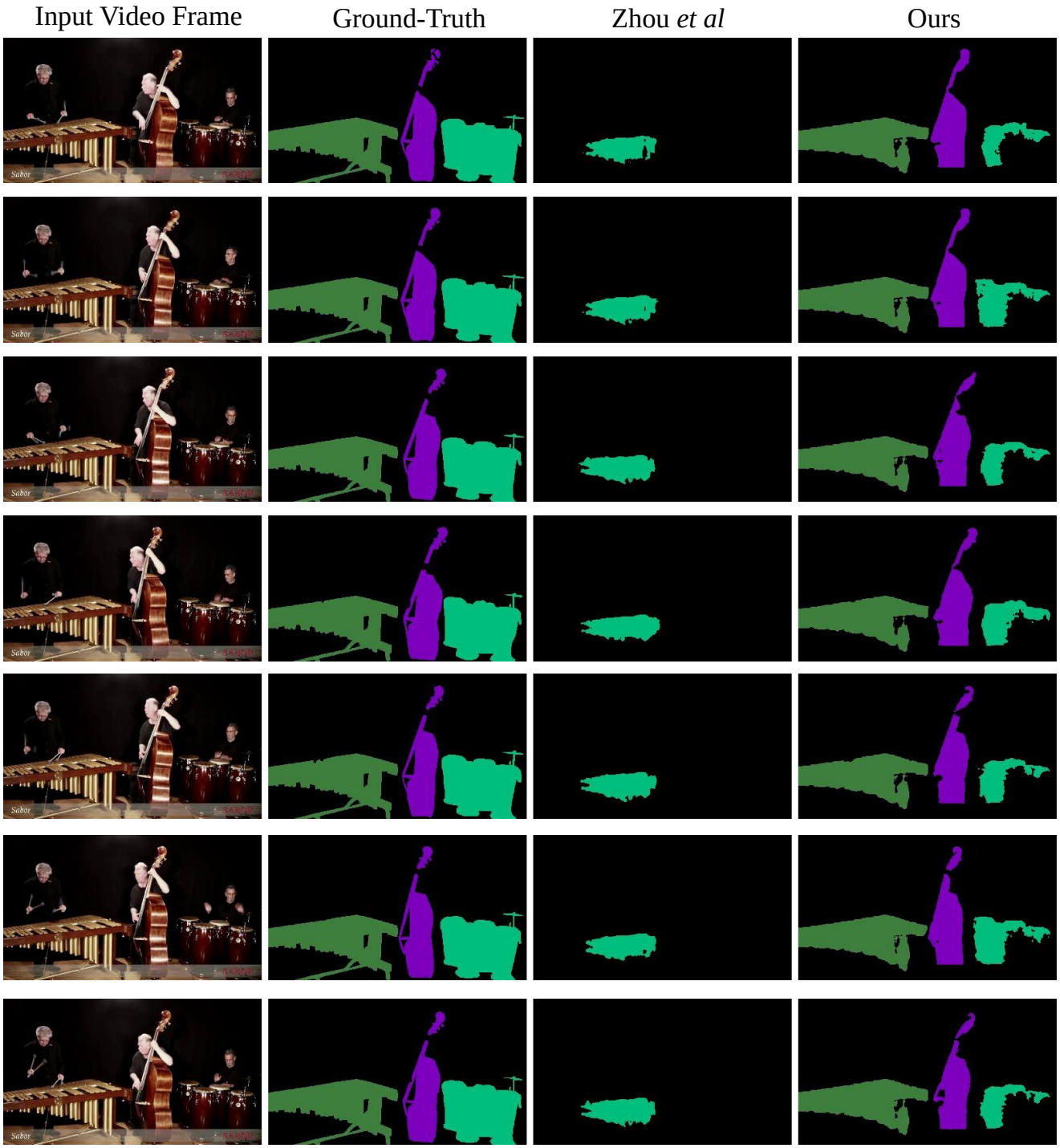

Figure E. Qualitative comparison to Zhou et al. [7] on AVS-Semantic. Each color represents a semantic category. Our model excels in accurately segmenting **multiple sounding objects**, showcasing its proficiency in audio-guided segmentation. This success can be attributed to the effective utilization of a decomposed and discretized audio representation, which enables the model to capture and analyze the distinct acoustic features of each object, resulting in precise segmentation outcomes.

eit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 3

[5] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao

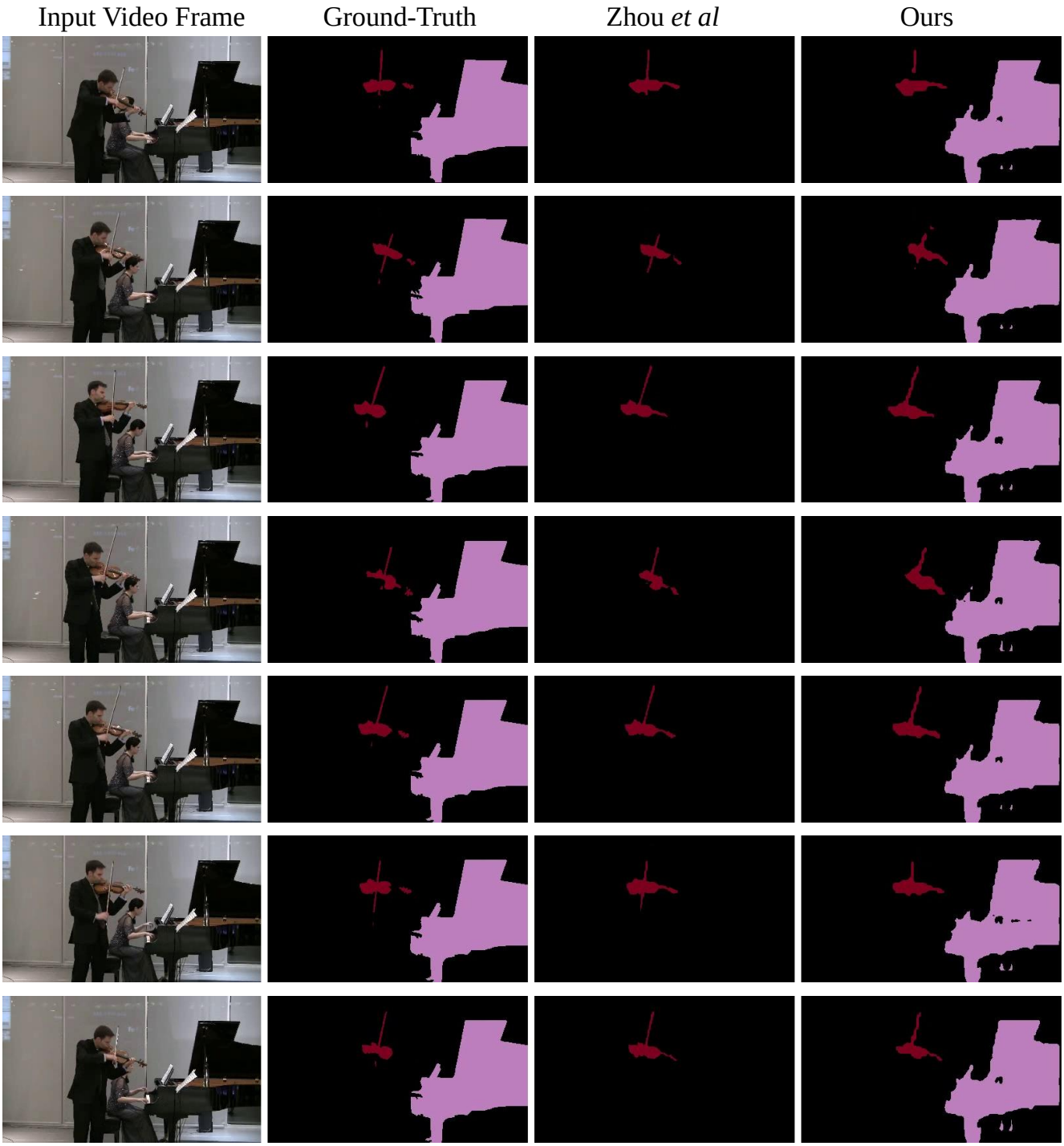

Figure F. Qualitative comparison to Zhou et al. [7] on AVS-Semantic. Each color represents a semantic category. Our model excels in accurately segmenting **multiple sounding objects**, showcasing its proficiency in audio-guided segmentation. This success can be attributed to the effective utilization of a decomposed and discretized audio representation, which enables the model to capture and analyze the distinct acoustic features of each object, resulting in precise segmentation outcomes.

Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 1

[6] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun,

Figure G. Qualitative comparison to Zhou et al. [7] on AVS-Semantic. Each color represents a semantic category. Our model demonstrates remarkable proficiency in accurately segmenting **tiny sounding objects**, owing to the implementation of a decomposed and discretized audio representation. By leveraging this technique, our model effectively captures the intricate acoustic details and nuances of these small-sized objects, resulting in precise and reliable segmentation outcomes.

Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, 2022. 1

[7] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang,

Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 2, 4, 5, 6, 7, 8