Supplementary Material: Cyclic Learning for Binaural Audio Generation and Localization

Zhaojian Li, Bin Zhao, and Yuan Yuan*
School of Artificial Intelligence, OPtics and ElectroNics, Northwestern Polytechnical University, China
zj_lee@mail.nwpu.edu.cn, {binzhao111, y.yuan1.ieee}@gmail.com

In this supplementary material, we provide more details and experimental results. Specifically, we further explain the process of the visual sounding object localization model and the object-aware upmix model in Section 1. We provide an additional baseline for binaural generation and localization and introduce the equations of evaluation metrics in Section 2. Furthermore, the audiovisual correlation analysis and more quantitative and qualitative results of the baseline model and our approach are demonstrated in Section 3. In qualitative results, we provide more visualized examples, including differential spectrograms, waveform envelopes, and sounding object localization. Finally, the generated binaural audio of our approach and the compared methods can be viewed in the video in supplementary materials.

1. Details of Localization and Upmix Models

In this section, we provide more details about the proposed localization model and upmix model. Firstly, we introduce the Audio Network and Visual Network in the localization model. Then,we explain the object-scene awareness module and the semantic-spatial mining module. Finally, we describe the upmix model.

Audio Network. We employ the short-time Fourier transform (STFT) [4] to convert the raw mono waveform signal into a Mel spectrogram. Then, we utilize a pre-trained VGGish [3] to extract spectrogram features and serve as input to the visual sounding object localization model. During the training phase, the model's weights are frozen.

Visual Network. We utilize the pre-trained ResNet-18 [2] to extract the visual features of the center frame of the video chunk. To extract specific features, the visual encoders of the generation and localization models are separated. The weights of the visual network are not frozen to obtain fine-tuned visual features for localization and generation.

Object-Scene Awareness Module. The extracted spectrogram features are input into two fully connected layers and a normalization layer to obtain the final spectrogram features f_a^m. Taking positive visual features as an example, the same goes for negative visual features. The extracted positive visual features are fed into a normalization layer after spatial aggregation, obtaining the final positive visual features f_{nv}. Then, the visual object features and the pseudo-visual object features are obtained through Eq. (5) and Eq. (6) in the main text. Finally, the visual object features are input into two fully connected layers and resized to meet the needs of interaction with the final spectrogram features.

Semantic-Spatial Mining Module. The process for the SSM module and the OSA module is the same. The only difference between them is that the number of input and output nodes in the fully connected layer of SSM is twice that of OSA because the input is binaural audio.

Object-Aware Upmix Model. The upmixing model and the localization model are integrated into a framework for unified training. The rest network structure of the upmix model generally follows [5], which consists of 3 residual blocks, each block has m=10 dilated convolutional layers [7]. In the inferencing, mono audio of length L and the corresponding image sequence are intercepted to obtain visual sounding objects by the localization model. The sounding objects are injected into the upmix model to provide object-level guidance information. Then, sample x_T from $p(x_T) \sim \mathcal{N}(0, \mathbf{I})$ and run the fusion and reverse procedure to obtain a clean differential audio. Finally, binaural audio is generated through Eq. (3) in the main text.

2. Details of Additional Baseline and Metrics

In this section, we provide an additional baseline for binaural localization and generation. In addition, we introduce the evaluation metrics of this paper in detail, which provide a comprehensive evaluation of binaural audio.

Additional Baseline. A randomly sampled positive audio-visual pair is $\{V_i, A_{mi}\}$, and then a negative audiovisual pair is regarded as $\{V_i, A_{mj}\}$, where A_{mi} represents the i-th mono audio, $i \neq j$. Next, pre-trained ResNet-18 [2] is used to extract visual features f_v. At the same time, pre-trained VGGish [3] is used to extract positive audio features f_{pa} and negative audio features f_{na}, respectively. The vi-

*Corresponding author

sual object feature f'_v can be obtained through:

$$f'_v = f_v \cdot \sigma\left((f_v)^\top \cdot f_{pa}\right). \quad (1)$$

Then, the distance between positive and negative sample pairs is obtained by

$$(d_+, d_-) = (\|f'_v - f_{pa}\|_2, \|f'_v - f_{na}\|_2). \quad (2)$$

Finally, the baseline model can be optimized by

$$l = \|(D_+, D_-) - (0, 1)\|_2. \quad (3)$$

In the main text, six evaluation metrics are used to comprehensively evaluate binaural audio generated by different methods, including STFT Distance [1], Envelope (ENV) Distance [6], Wave L2 (WAV) [9], Amplitude L2 (AMP), Phase L2 (PHA), and Signal-to-Noise Ratio (SNR) [8].

STFT Distance: It measures binaural audio on the spectrogram domain, which is the Euclidean distance between the predicted left and right channel spectrograms and their ground-truth:

$$\mathcal{D}_S = \|S_b^l - \hat{S}_b^l\|_2 + \|S_b^r - \hat{S}_b^r\|_2. \quad (4)$$

ENV Distance: It measures binaural audio on the raw waveform domain, which is the Euclidean distance between the envelope of the predicted waveform's left and right channels and its ground-truth:

$$\mathcal{D}_E = \|E[A_b^l] - E[\hat{A}_b^l]\|_2 + \|E[A_b^r] - E[\hat{A}_b^r]\|_2, \quad (5)$$

where $E[\cdot]$ denote the envelope of signal. It can capture the perceptual similarity of the waveform well.

Wave L2: It is the mean squared error between the predicted binaural audio and the ground-truth binaural recording.

$$\mathcal{L}_2^{wav} = (A_b^l - \hat{A}_b^l)^2 + (A_b^r - \hat{A}_b^r)^2. \quad (6)$$

Amplitude L2 and Phase L2: Amplitude L2 and Phase L2 are the mean squared errors between the predicted binaural audio and the real binaural recording on the amplitude and phase after STFT on the waveform:

$$\mathcal{L}_2^{amp} = (|S_b^l| - |\hat{S}_b^l|)^2 + (|S_b^r| - |\hat{S}_b^r|)^2, \quad (7)$$

and

$$\mathcal{L}_2^{pha} = (\angle(S_b^l) - \angle(\hat{S}_b^l))^2 + (\angle(S_b^r) - \angle(\hat{S}_b^r))^2, \quad (8)$$

where $|\cdot|$ and $\angle(\cdot)$ denote the modulu and phase angle of the complex number, respectively.

Signal-to-Noise Ratio: It is the power ratio of a signal to noise. Signal refers to the ground-truth binaural recording, while noise refers to the differential between the ground truth and the predicted signal.

$$\text{SNR} = \frac{10}{2} \cdot (\log 10(\frac{A_b^l}{A_b^l - \hat{A}_b^l}) + \log 10(\frac{A_b^r}{A_b^r - \hat{A}_b^r})). \quad (9)$$

Method	Distance↓	Accuracy(%)↑
Additional Baseline	0.083	92.8
Ours	0.028	96.1

Table 1. Quantitative results of baseline model on sounding object localization.

Method	STFT↓	ENV↓	WAV↓
Additional Baseline	0.799	0.129	5.328
Ours	0.779	0.128	5.200

Table 2. Quantitative results of baseline model on binaural audio generation.

We evaluate the performance of the localization model using audiovisual distance and classification accuracy on the entire testset. Note: We counted the results of all 0.1s sliding windows for each 10s clip in the testset. We simultaneously sample positive audiovisual pairs and negative audiovisual pairs to compute the audiovisual distance between them. We employ a softmax function to scale the distance between the audio and visual features to 1.0. Ideally, the distance D_+ of positive audiovisual pairs should tend to 0.0, while the distance D_- of negative audiovisual pairs should tend to 1.0:

$$\text{distance} = \|(D_+, D_-) - (0, 1)\|_2. \quad (10)$$

When the audiovisual distance ≤ 0.5, we determine that the audiovisual pair predicted by the model is paired. Otherwise, it is unpaired. Then, the classification accuracy of the localization model can be expressed as:

$$\text{accuracy} = \frac{TP + TN}{2 \times (TP + FN)}, \quad (11)$$

where TP and TN refer to the number of correctly predicted pairs and non-pairs, respectively. FN represents the number of pairs predicted to be non-pairs.

3. Audiovisual Correlation Analysis and More Quantitative and Qualitative Results

In this section, we first present the quantitative and qualitative results of the additional baseline. Then, we analyze the correlation between audio and visual modalities under the baseline model and our approach. Finally, more visualization results are employed to demonstrate the superiority of our approach.

Baseline Model Results. Table 1 shows the comparison results between the additional baseline and our method on sounding object localization. It can be seen that our method outperforms the additional baseline in both audiovisual distance and classification accuracy. In Fig. 5, the first column shows the qualitative results of the additional baseline

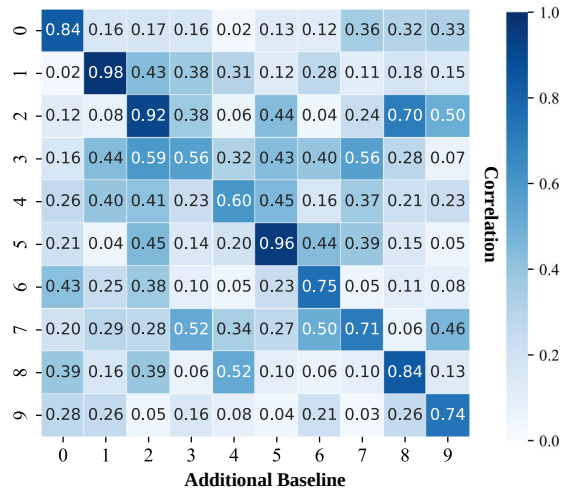

Figure 1. Audiovisual correlation analysis of baseline model.

model. Compared to the additional baseline, our method achieves better localization performance. Further, we combine additional baseline with our upmix model to generate binaural audio. The generation results of the additional baseline are shown in Table 2. It can be seen that our method outperforms the baseline model in binaural audio generation. This demonstrates that good localization results can improve binaural audio generation performance.

Audiovisual Correlation Analysis. To intuitively observe the ability of different losses and modules to correlate objects and sounds, we performed an audiovisual correlation analysis. Specifically, we randomly select 10 positive audiovisual samples from the testset, where each positive sample corresponds to 9 negative audiovisual samples. We obtained a total of 100 audiovisual sample pairs. Then, we compute and visualize the distance between audio and visual modalities, as shown in Fig. 1 and Fig. 2. When i=j, the audio and visual samples are positive pairs, and the rest are negative samples. Fig. 1 shows the correlation analysis results of the baseline model. It can be seen that the correlation of positive audiovisual samples output by the baseline model is greater than 50%. However, its confidence level is not high. Fig. 2 shows the correlation analysis results of our method under different losses. Compared to the baseline model, our method can better correlate sounds and objects and be more discriminative (see Fig. 2 (a) and Fig. 2 (b)) . In addition, it can be seen from Fig. 2 (c) that the combination of l_{loc}^{sce} and l_{loc}^{obj} can further enhance the correlation of positive audiovisual samples and suppress the correlation of negative audiovisual samples.

More Qualitative Results. We present more qualitative results in Fig. 3, Fig. 4, and Fig. 5. Fig. 3 and Fig. 4 show the visualization results of our method and other methods on spectrogram and waveform, respectively. Fig. 5 shows the visualization results of additional baseline, proposed losses, and modules on sounding object localization. Overall, compared with other methods, our method shows superior binaural generation and localization performance.

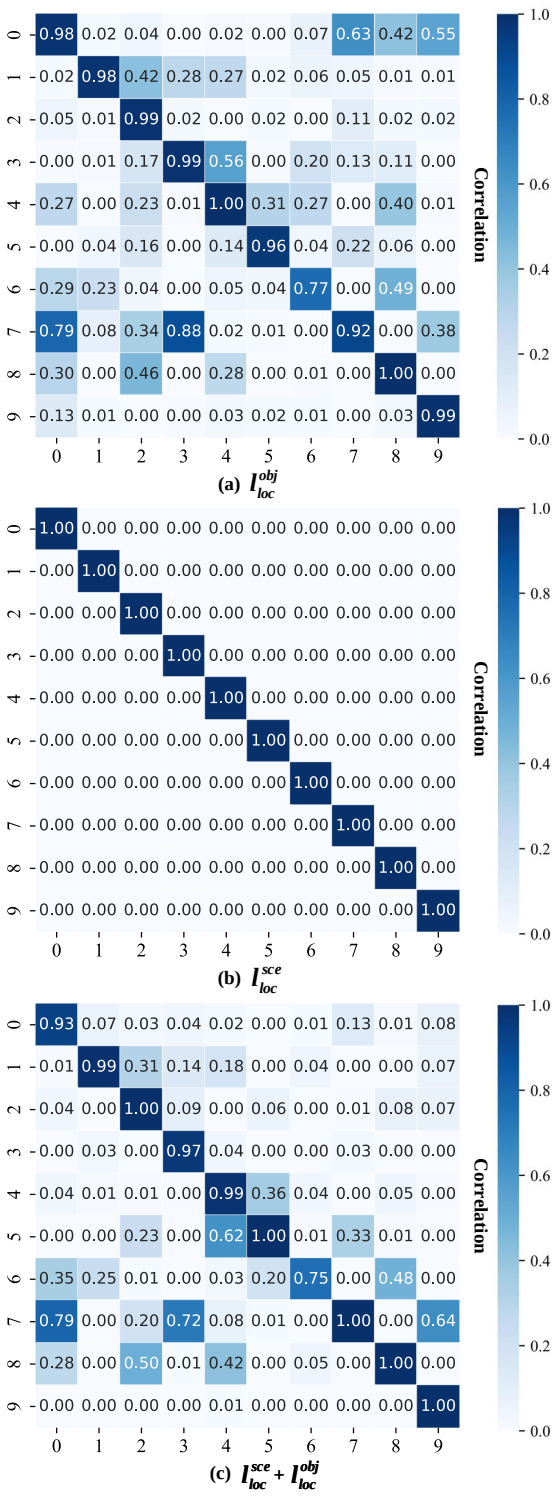

Figure 2. Audiovisual correlation analysis of our approach.

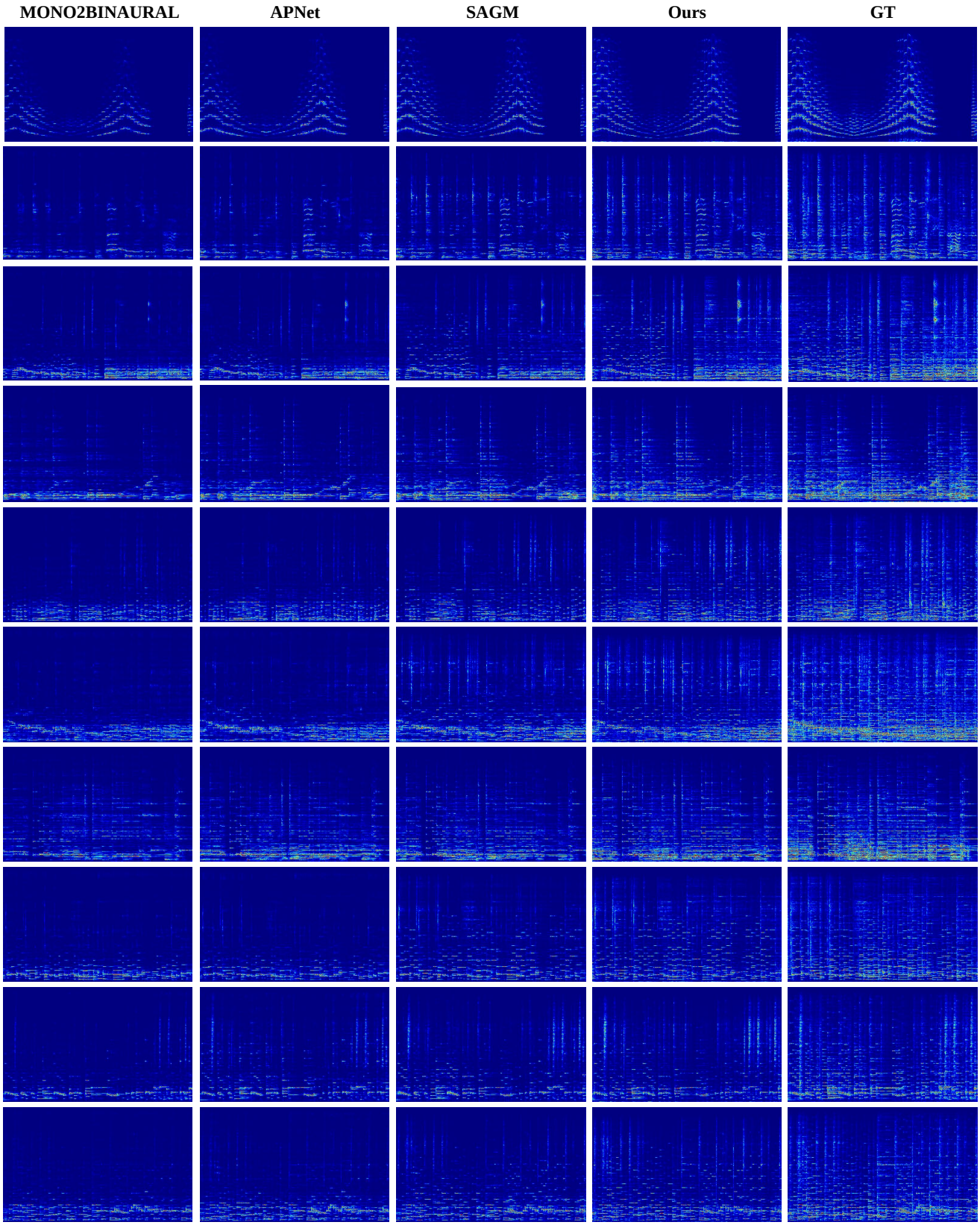

Figure 3. More qualitative results for audio differential spectrograms.

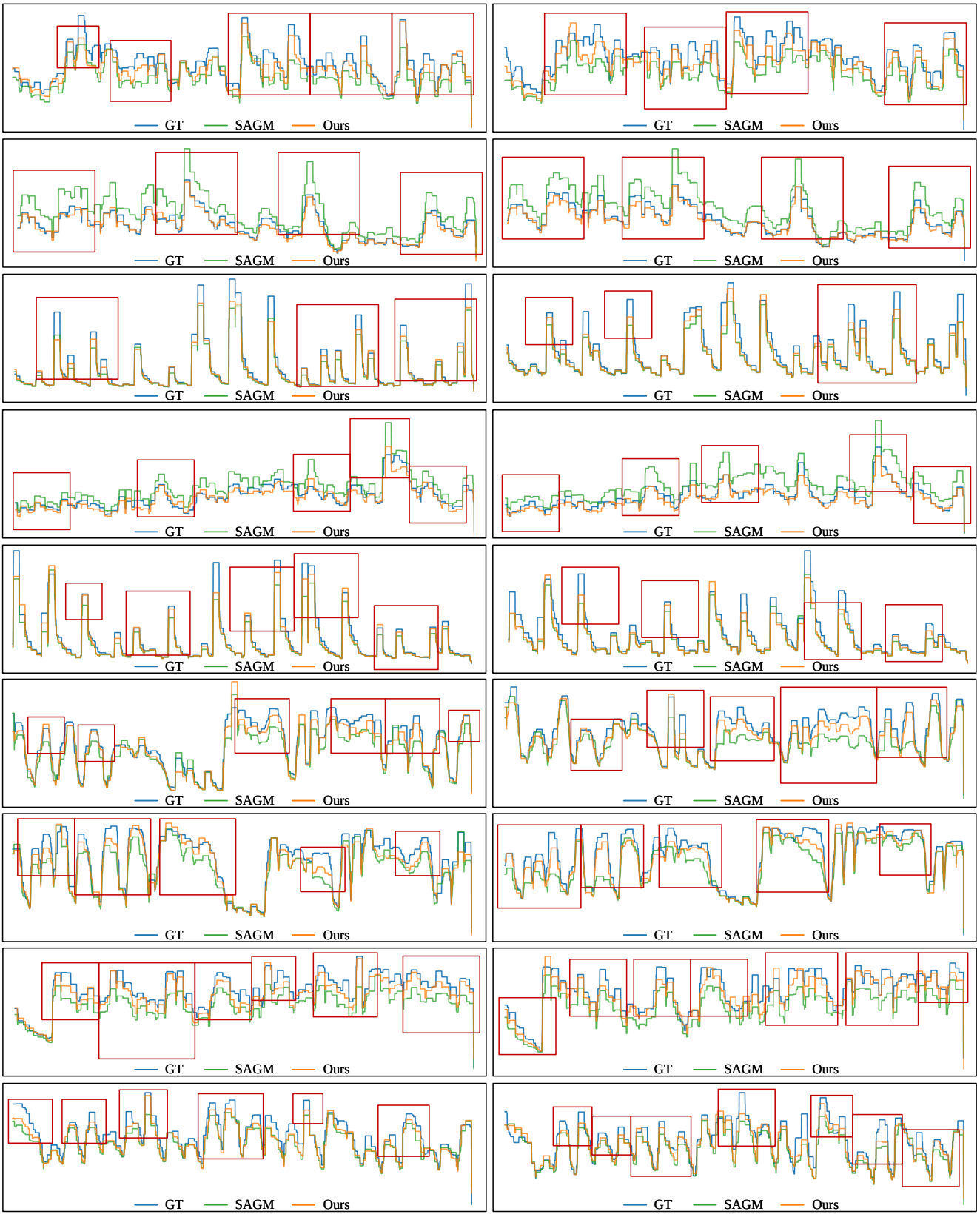

Figure 4. More qualitative results of binaural waveform envelopes. The left and right columns represent the left and right channels, respectively.

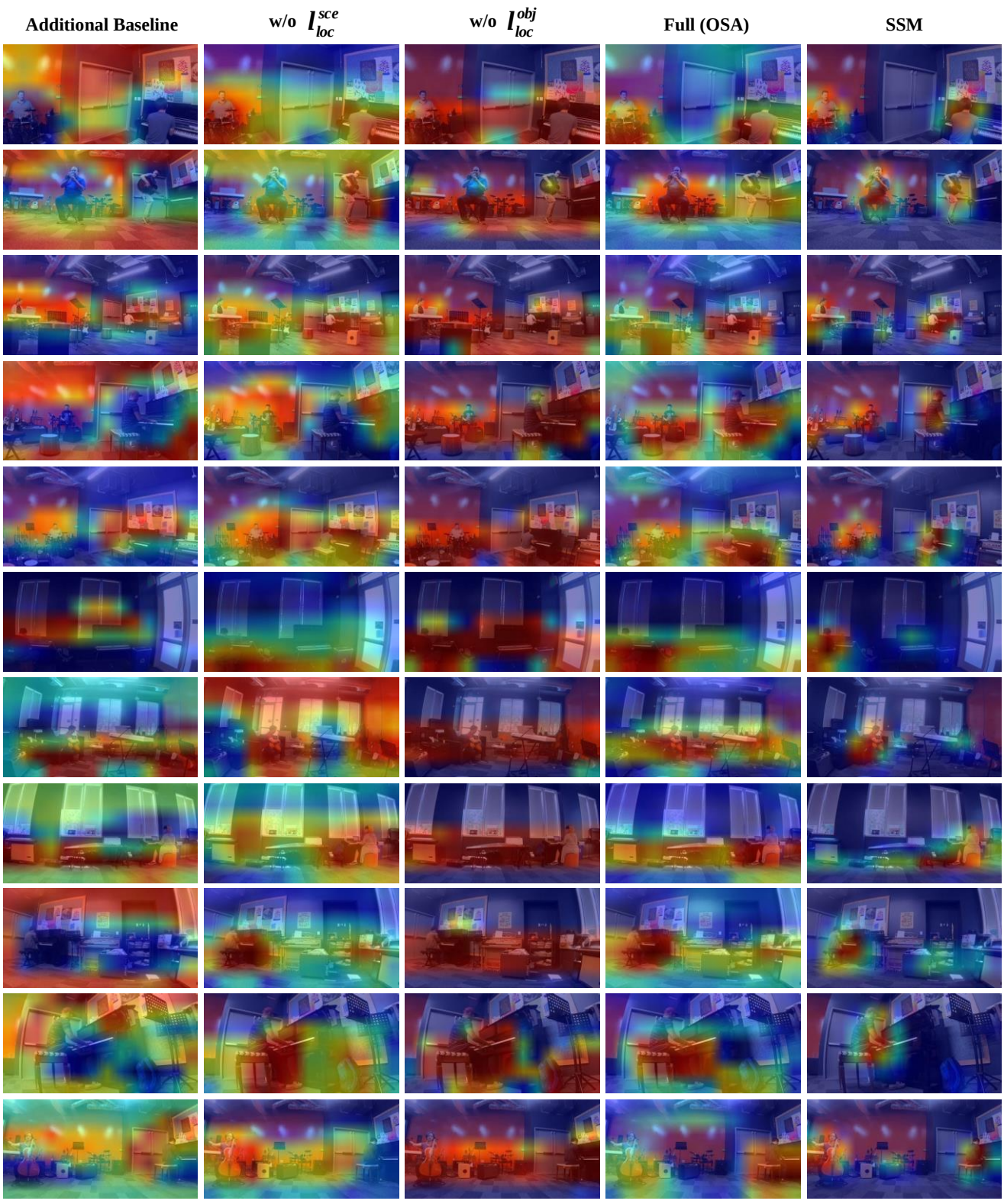

Figure 5. More qualitative results of the sounding object localization model.

References

[1] Ruohan Gao and Kristen Grauman. 2.5D Visual Sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[3] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 131–135, 2017. 1

[4] Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. iSTFTNet: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6207–6211, 2022. 1

[5] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 1

[6] Yan-Bo Lin and Yu-Chiang Frank Wang. Exploiting Audio-Visual Consistency with Partial Supervision for Spatial Audio Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2056–2063, 2021. 2

[7] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499*, 2016. 1

[8] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond Mono to Binaural: Generating Binaural Audio from Mono Audio with Depth and Cross Modal Attention. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 3347–3356, 2022. 2

[9] Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando De la Torre, and Yaser Sheikh. Neural Synthesis of Binaural Speech From Mono Audio. In *Proceedings of the International Conference on Learning Representations*, 2021. 2