– Supplementary Material –
How Far Can We Go With Synthetic Data for Audio-Visual Sound Source Localization?

The contents in this supplementary material are as follows:

Contents

1. Datasets

We use the following datasets for evaluation, following [12]:

VGGSS [1]. It includes bounding box annotations for ∼5K samples and is used in the Single Sound Source Localization task, where these annotations are incorporated for evaluation. Both image and audio samples in this dataset are real.

IS4 [12]. This dataset consists of 3240 images, generating 6480 unique audio-visual instances (with two objects per image), where each object is paired with a corresponding audio. It is used for both Single Sound Source Localization and Interactive Localization tasks. In Single Sound Source Localization, each unique pair is evaluated independently, whereas in Interactive Localization, both pairs are considered together. The images in this dataset are synthetically generated and paired with real audio samples. It provides both segmentation and bounding box annotations.

VPO Benchmark [14]. VPO consists of two datasets: VPO-SS (Single Source) and VPO-MS (Multi-Source), containing 890 and 1,437 samples, respectively. The images are sourced from the COCO dataset and paired with audio samples from VGGSound. Since the images come from COCO, they include segmentation mask annotations. These datasets are used in Audio-Visual Segmentation, Sound Source Localization, and Interactive Localization tasks, with VPO-MS used exclusively for the latter.

AVSBench [15]. AVSBench consists of two datasets: AVS-S4 (Single Source), which contains a single sound-producing object, and AVS-MS3 (Multi-Source), where multiple objects generate sound simultaneously. We use these datasets for the Audio-Visual Segmentation task, as they were originally designed for, and additionally use AVS-S4 for Single Sound Source Localization.

AVSBench-Robust S4 [7]. This dataset is used to evaluate the Audio-Visual Robustness task introduced in [7]. We present the results of this task in Section 3. AVSBench-Robust S4 is an extension of the AVS-S4 dataset with three additional negative audio conditions alongside the original positive audio. The negative conditions include Silence (no audio but silent visual frames), Noise (*e.g.*, white noise), and Off-screen Audio (semantically unrelated sounds from different categories). The dataset contains 4,932 samples: 3,452 for training, 740 for validation, and 740 for testing.

We note that several datasets are partially derived from VGGSound by reusing its audio, such as IS4 and the VPO variants. To prevent any train–test leakage, we carefully checked for overlaps and removed all conflicting samples. In IS4, the image side consists of text-to-image generations, while the audio is taken from VGGSound, and the two are synthetically paired. In VPO, the images come from COCO and the audio from VGGSound, again synthetically paired. By contrast, VGGSS and AVSBench are naturally paired, in-the-wild real-world datasets.

2. Details on Evaluation Metrics

cIoU and AUC [10]. These metrics are widely adopted for evaluating sound source localization. cIoU measures the spatial correspondence between the predicted sounding region and the annotated bounding box, whereas AUC summarizes performance across varying cIoU thresholds. Under the conventional evaluation protocol, the predicted region is obtained by selecting the top 50% of pixels with the highest activations in the audio-visual attention map. The Intersection over Union (IoU) between this region and the ground-truth box is then computed, and a prediction is considered correct when the IoU exceeds 0.5.

cIoU Adaptive and AUC Adaptive [12]. A limitation of the conventional cIoU protocol is that it selects the top 50% of pixels in the attention map as the predicted sounding region, a heuristic that disproportionately penalizes cases with small sound sources. Because the predicted region is fixed to occupy half of the image, over-localization becomes inevitable when the true source is much smaller, leading even accurate models to obtain low IoU scores. To address this issue, [12] introduced modified metrics that determine the number of relevant pixels, K, directly from the ground-truth annotation and evaluate localization using only the top-K activated pixels, thereby aligning the evaluation mask size with the actual object extent.

mIoU and F-Score [15]. mIoU and F-Score are conventional metrics for evaluating segmentation performance and are directly adopted in audio-visual segmentation tasks [15].

mIoU Adaptive and F-Score Adaptive [12]. Following [12], adaptive variants of the segmentation metrics are employed. Their formulation follows the same principle described above, extending the same GT pixel-count–based evaluation protocol to segmentation.

IIoU and IAUC [12]. To assess interactive localization, where a single image is paired with multiple audio signals corresponding to different sound sources in the image, IIoU and IAUC were introduced in [12]. These metrics extend cIoU and Adaptive cIoU to evaluate localization accuracy for each audio–same image pair. A sample is regarded as correct under IIoU only when all sound sources are localized accurately; mislocalization of any source results in the entire sample being counted as a failure.

IIoU Adaptive and IAUC Adaptive [12]. The adaptive variants apply the same GT pixel-count–based evaluation scheme to IIoU and IAUC, following the adaptive formulation described above.

G-mIoU, G-F, and G-FPR [7]. These metrics are used for the Audio-Visual Robustness task, for which we present the results in Section 3 in this material, and are directly adopted from [7]. Introduced in [7], the Global mIoU (G-mIoU), G-F score, and G-FPR evaluate audio-visual robustness against non-corresponding audio (*e.g.*, silence or off-screen sounds). G-mIoU and G-F globally aggregate performance across positive cases (accurate segmentation when the sound source is visible) and negative cases (where the ideal output is an empty mask). The Global False Positive Rate (G-FPR) averages false activations over negative samples, reflecting how often the model incorrectly predicts a mask when none should exist.

Method	G-mIoU (↑)	G-F (↑)	G-FPR (↓)
1x Scaled Dataset:			
Original	71.226 $\pm$ 2.578	69.823 $\pm$ 1.821	0.022 $\pm$ 0.013
Synthetic	70.367 $\pm$ 0.451	68.724 $\pm$ 0.382	0.053 $\pm$ 0.032
(SynI,RealA)	74.679 $\pm$ 1.083	72.068 $\pm$ 0.877	0.017 $\pm$ 0.006
(RealI,SynA)	67.031 $\pm$ 2.308	66.606 $\pm$ 1.594	0.032 $\pm$ 0.016
(SynI,MixedA)	72.046 $\pm$ 1.012	69.852 $\pm$ 0.865	0.040 $\pm$ 0.018
(MixedI,SynA)	71.680 $\pm$ 1.862	69.561 $\pm$ 1.582	0.041 $\pm$ 0.018
2x Scaled Dataset:			
(A) {(S, S) ∪ (S, S)}	68.686 $\pm$ 2.255	67.480 $\pm$ 1.840	0.052 $\pm$ 0.031
(B) {(S, R) ∪ (S, R)}	73.664 $\pm$ 1.224	71.276 $\pm$ 1.031	0.018 $\pm$ 0.008
(C) {(R, R) ∪ (S, R)}	76.146 $\pm$ 0.713	73.439 $\pm$ 0.545	0.017 $\pm$ 0.006
(D) {(R, R) ∪ (S, S)}	75.278 $\pm$ 0.835	72.721 $\pm$ 0.568	0.028 $\pm$ 0.007
(E) {(S, R) ∪ (S, S)}	74.298 $\pm$ 1.084	71.862 $\pm$ 0.822	0.017 $\pm$ 0.011
3x Scaled Dataset:			
(A) {(S, S) ∪ (S, S) ∪ (S, S)}	70.797 $\pm$ 2.085	69.265 $\pm$ 1.463	0.031 $\pm$ 0.008
(B) {(S, R) ∪ (S, R) ∪ (S, R)}	73.965 $\pm$ 0.333	71.625 $\pm$ 0.247	0.013 $\pm$ 0.005
(C) {(R, R) ∪ (S, R) ∪ (S, R)}	76.134 $\pm$ 0.904	73.369 $\pm$ 0.704	0.008 $\pm$ 0.002

Table 1. **Performance comparison on AVSBench-Robust S4.** *Mixed* denotes both synthetic and real samples within the same modality are used in a mixed form as (134K Syn. + 10K Real). **VGGSound** (R: 144K Real Images, R: 144K Real Audios), **VGGSyn1** (S: 144K Synthetic Images, S: 144K Synthetic Audios), **VGGSyn2** (S: 144K Synthetic Images, S: 144K Synthetic Audios) and **VGGSyn3** (S: 144K Synthetic Images, S: 144K Synthetic Audios).

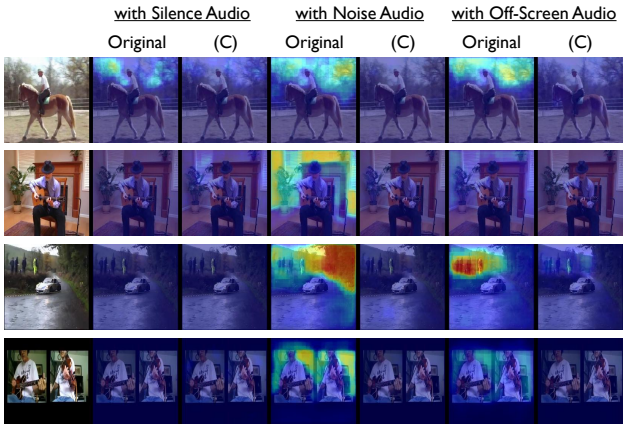

Figure 1. **AVSBench-Robust S4 results.** Original refers to the 1× configuration using only real image–audio pairs, and the 3× result corresponds to configuration (C) {(R, R) ∪ (S, R) ∪ (S, R)}; both configurations are defined in Table 1.

Method	Localization				Segmentation				Interactive			
	cIoU	w/ Adap.	AUC	w/ Adap.	mIoU	w/ Adap.	F-Score	w/ Adap.	IIoU	w/ Adap.	IAUC	w/ Adap.
Original	48.03	62.22	41.95	51.99	47.56	51.85	53.70	60.26	31.10	45.71	29.47	40.40
Original-Corr. [11, 13]	49.97	62.83	42.54	52.23	47.88	52.63	54.45	61.29	31.09	46.76	29.47	40.36
Original-ImageBind [5]	50.14	63.43	43.21	52.77	48.95	52.76	54.92	61.10	33.43	48.58	30.73	41.44
(SynI,RealA)	55.13	67.16	46.73	55.06	50.78	53.57	56.47	62.01	41.49	56.61	37.01	46.69

Table 2. **Comparison of Frame-Selection Methods.**

Method	Localization				Segmentation				Interactive			
	cIoU	w/ Adap.	AUC	w/ Adap.	mIoU	w/ Adap.	F-Score	w/ Adap.	IIoU	w/ Adap.	IAUC	w/ Adap.
Original	48.03	62.22	41.95	51.99	47.56	51.85	53.70	60.26	31.10	45.71	29.47	40.40
Synthetic w/ Stable Diffusion 3 Medium [2]	47.97	60.88	41.86	51.03	48.10	51.92	53.24	60.25	31.38	46.44	29.95	40.24
Synthetic w/ Flux.1 [schnell] [6]	50.08	62.43	43.48	51.88	47.90	52.18	53.42	60.66	34.53	49.02	32.36	41.61
(SynI,RealA) w/ Stable Diffusion 3 Medium [2]	55.13	67.16	46.73	55.06	50.78	53.57	56.47	62.01	41.49	56.61	37.01	46.69
(SynI,RealA) w/ Flux.1 [schnell] [6]	55.25	67.23	46.97	55.05	50.64	54.03	56.88	62.72	41.97	56.69	37.52	46.49

Table 3. **Comparison between Stable Diffusion 3 Medium [2] and Flux.1 [schnell] [6].**

Method	Localization				Segmentation				Interactive			
	cIoU	w/ Adap.	AUC	w/ Adap.	mIoU	w/ Adap.	F-Score	w/ Adap.	IIoU	w/ Adap.	IAUC	w/ Adap.
Original	48.03	62.22	41.95	51.99	47.56	51.85	53.70	60.26	31.10	45.71	29.47	40.40
Synthetic w/ Stable Audio 1.0 [3]	47.97	60.88	41.86	51.03	48.10	51.92	53.24	60.25	31.38	46.44	29.95	40.24
Synthetic w/ Tango 2 [8]	42.32	56.09	39.14	47.97	41.93	47.51	47.54	55.80	24.78	39.85	26.67	36.33

Table 4. **Comparison between Stable Audio Open 1.0 [3] and Tango 2 [8].**

3. Audio-Visual Robustness Task

We present one additional task, Audio–Visual Robustness, on top of the main tasks provided in the main paper. This task covers scenarios in which the auditory and visual signals are intentionally mismatched, such as silent audio, noisy audio (*e.g.*, white noise), or off-screen sound, to evaluate whether models can reliably detect when a sound does not originate from a visible objects. Variants of this type of task have been introduced in [7, 9], and we follow [7] to assess our model's ability. The results of our model for the 1×, 2×, and 3× scales are presented in Table 1. As the results indicate, our method shows trends consistent with those observed in the main tasks of the paper, particularly in the 1× setting. Additionally, scaling up the data improves performance relative to the 1× scale. These findings demonstrate that our new data-centric strategy with synthetic data is beneficial and pushes model performance far beyond that of real-data-only training across a wide variety of sound source localization tasks. We also provide qualitative results for this task in Figure 1.

4. Performance Analysis with Non-Naive Frame Selections

We show in the main paper that mid-frame selection, the most widely used strategy for obtaining the visual side of audio–visual pairs, introduces imperfections and misalignment between audio and images. In contrast, synthetic images enhance semantic consistency when paired with real audio, proving highly beneficial throughout the paper. However, one may argue that mid-frame selection is simply a poor policy, making it difficult to assess the real gain achieved by synthetic data. To examine this possibility, we train the original model on real data using two alternative non-naive frame-selection strategies:

- **Correlation-based pairs:** We use images of highly correlated audio–visual pairs computed in [11, 13] from VGGSound videos, which were shown to be effective in those prior works.
- **ImageBind-guided selection:** We compute ImageBind [5] similarity scores for each audio segment and the corresponding visual frames in a video, and select the frame with the highest similarity as the representative frame for training.

Experimental results are provided in Table 2. Although both strategies slightly outperform mid-frame selection, they still fall short of our *(SynI, RealA)* variant with large margin, which uses synthetic images. Since all methods rely on the same real audio, this clearly demonstrates the effectiveness of synthetic images in overcoming imperfections and semantic misalignment.

5. Ablation with Other Text-to-X Models

In Section 4.1 of the main paper, we state that although stronger and more specialized models could be used, we aim to rely on standard and widely available generative models to establish a more cost-effective recipe. To this end, we analyze the impact of alternative generative models when used in place of our standard choices.

We first examine the image side. We replace our standard Stable Diffusion 3 Medium [2] with a more recent but computationally heavier model Flux.1 [schnell] [6] (12 GB vs 33 GB respectively). All settings remain the same, except synthetic images are generated with Flux.1 [schnell], and the data scale is limited to 1× due to computational

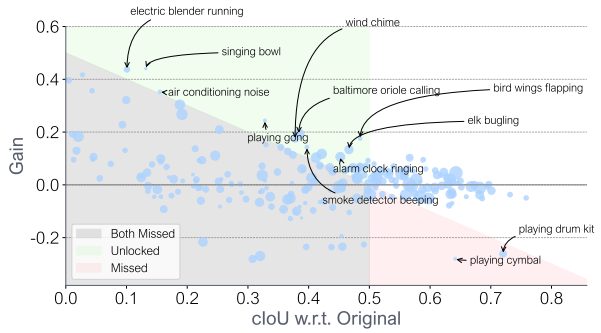

Figure 2. **Class-wise performance analysis: Original vs. (SynI, RealA).** The circle size represents the number of samples in each category.

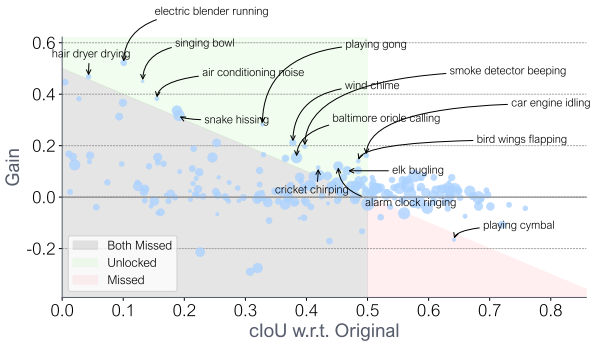

Figure 3. **Class-wise performance analysis: Original vs. {(R, R) ∪ (S, R) ∪ (S, R)}.** The circle size represents the number of samples in each category.

constraints. The experimental results are shown in Table 3. As the results indicate, in the fully synthetic setup, Flux.1 [schnell] yields additional improvements, especially on the Localization and Interactive tasks. However, in the *(SynI, RealA)* setup, this performance gap largely disappears compared to Stable Diffusion 3 Medium. Nevertheless, if one intends to rely solely on synthetic data – either due to the absence of real data or the impossibility of collecting new data – the more expensive recipe 💰 "using Flux.1 [schnell] for image generation and Stable Audio Open 1.0 [3] for audio generation in the fully synthetic mode" is beneficial and can even surpass training with real data.

Second, we explore the audio side. We replace Stable Audio Open 1.0 [3] with another popular audio generation model, Tango 2 [8]. All settings remain the same, except synthetic audio is generated using Tango 2, and the data scale is limited to 1×. The experimental results are shown in Table 4. As the results indicate, in the fully synthetic setup, Tango 2 yields substantially lower performance compared to our design choice, Stable Audio Open 1.0. This outcome is expected, as we observe that Tango 2 fails to generate audio for many class categories or concepts in VGGSound. Overall, this comparison further validates the use of Stable Audio Open 1.0 in our recipe for synthetic audio generation.

6. Class-wise Performance Analysis on VGGSS

Although class labels are never used during training in any of our experiments, the VGGSS dataset contains class label information. This allows us to conduct a Class-wise Performance Analysis to determine which categories benefit from our proposed variants and which struggle. Additionally, we can identify newly unlocked categories enabled by synthetic data. We present this analysis in Figure 2 and Figure 3 for the variants *(SynI, RealA)*, where VGGSound images are replaced with synthetic ones, and for variant {(R, R) ∪ (S, R) ∪ (S, R)}, the final model that achieves state-of-the-art performance. It is important to note that while some categories may appear successful or unsuccessful in this analysis due to aggregated results, individual samples may still achieve high or low accuracy and contribute to the final performance.

Figure 2 illustrates which categories benefit from replacing real images with synthetic ones. As shown in the green zone, this approach unlocks new categories, such as 'electric blender running' and 'elk bugling', which were previously not well localized by the original model. While some categories remain in the gray zone, the majority show a positive gain, and individual samples within these categories contribute significantly to the final performance.

Figure 3 presents the analysis for variant {(R, R) ∪ (S, R) ∪ (S, R)}, the state-of-the-art model in our study. Similar to *(SynI, RealA)*, it unlocks new categories beyond those introduced by *(SynI, RealA)*, such as 'snake hissing', 'cricket chirping', and 'hair dryer drying'. Additionally, the positive side of the gray zone appears more populated, which directly correlates with the final state-of-the-art performance. An interesting observation is that 'playing cymbal' consistently appears as a failed category. To investigate whether this results from confusing images, we visualize its synthetic images. Similarly, we also compared images from some newly unlocked categories in Figure 4. A comparison of images in the 'playing cymbal' category reveals that our approach tends to generate a single cymbal plate, whereas real samples typically depict a cymbal within a drum setup. This difference in domain and style leads to a performance drop in this category. In contrast, when examining newly unlocked categories, we observe that synthetic images provide clearer semantics. For instance, objects of interest – such as crickets, elks, and snakes – are consistently visible in synthetic images, whereas their real counterparts often lack these elements due to mid-frame selection.

7. Visualization of Generated Samples

In this section, we present examples of synthetic data generated using our proposed pipeline. In Figure 5, we visualize the generated image samples. These images not only accurately contain the given semantic objects or actions but also enhance diversity by placing them in unusual environments, such as a 'racing car on an ice rink', 'robots playing table tennis on the moon', or 'a person playing guitar on a distant planet's surface'. In Figure 6, we present examples of generated audio samples from our synthetic clones of VGGSound. For visualization purposes, since audio cannot be played, we use an off-the-shelf audio captioning model [4] to describe the audio samples. As shown, the generated audio aligns with the intended class categories.

8. Detailed Results for Datasets and Tasks

In Section 4.2 and 4.3 (main paper), we present results for each dataset and task using bar graphs. Here, we provide detailed numerical results in tables. Table 5, Table 6, and Table 7 show the results for 1×, 2×, and 3× scales, respectively.

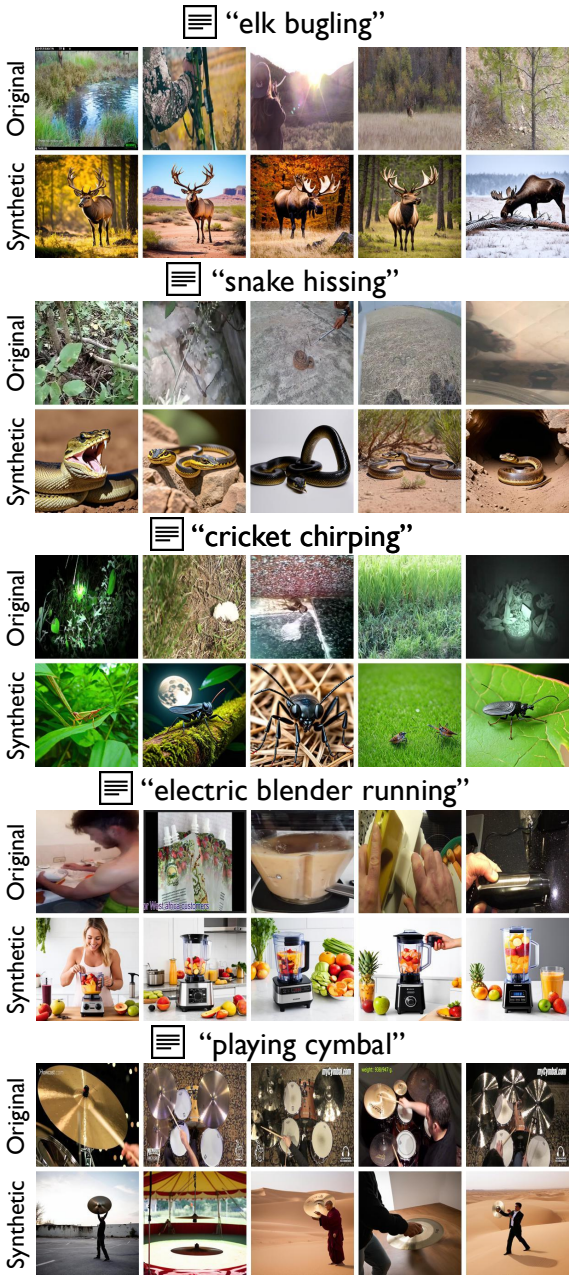

Figure 4. **Comparison of Original Images vs. Synthetic Images**

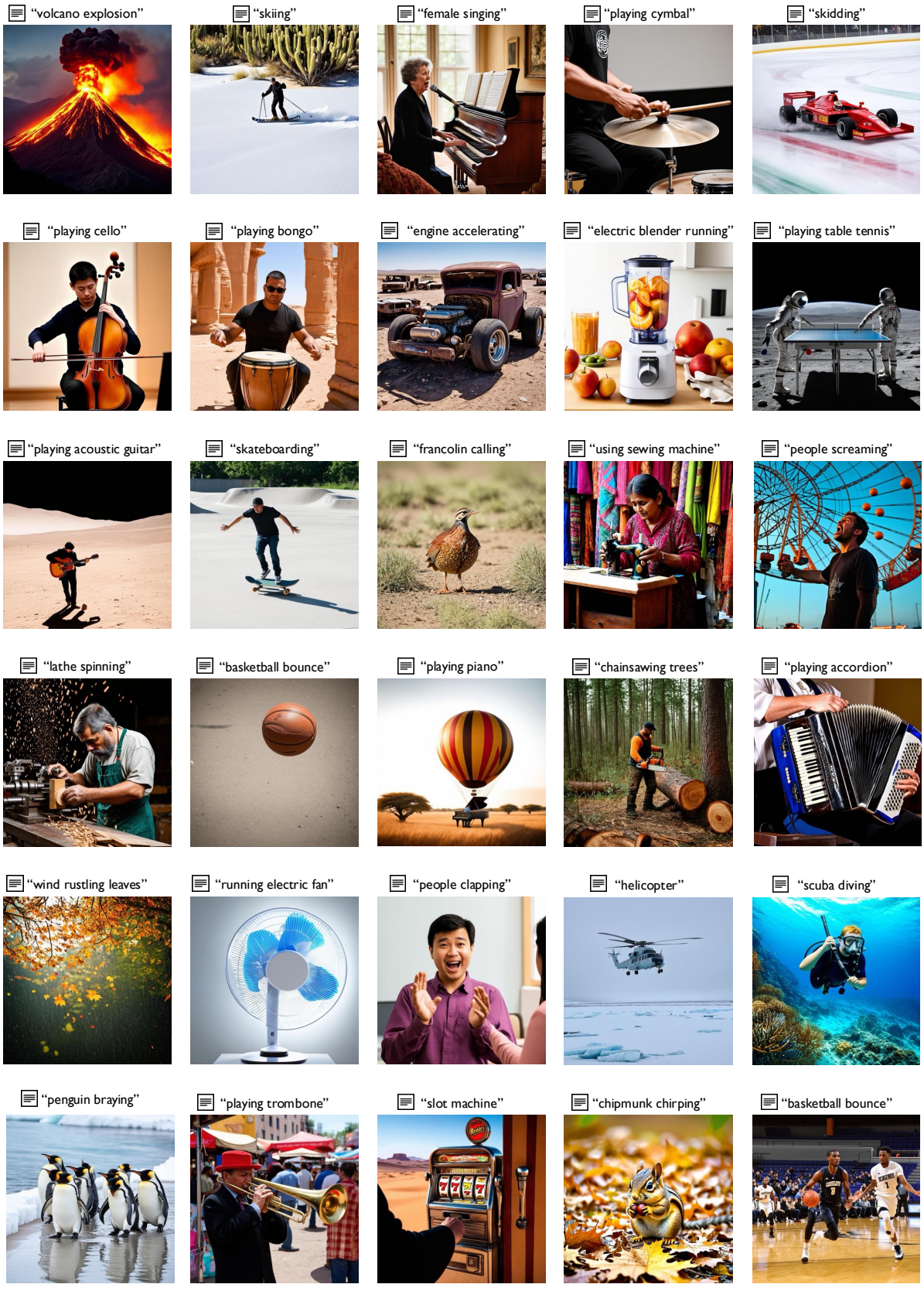

Figure 5. **Generated Image Examples.**

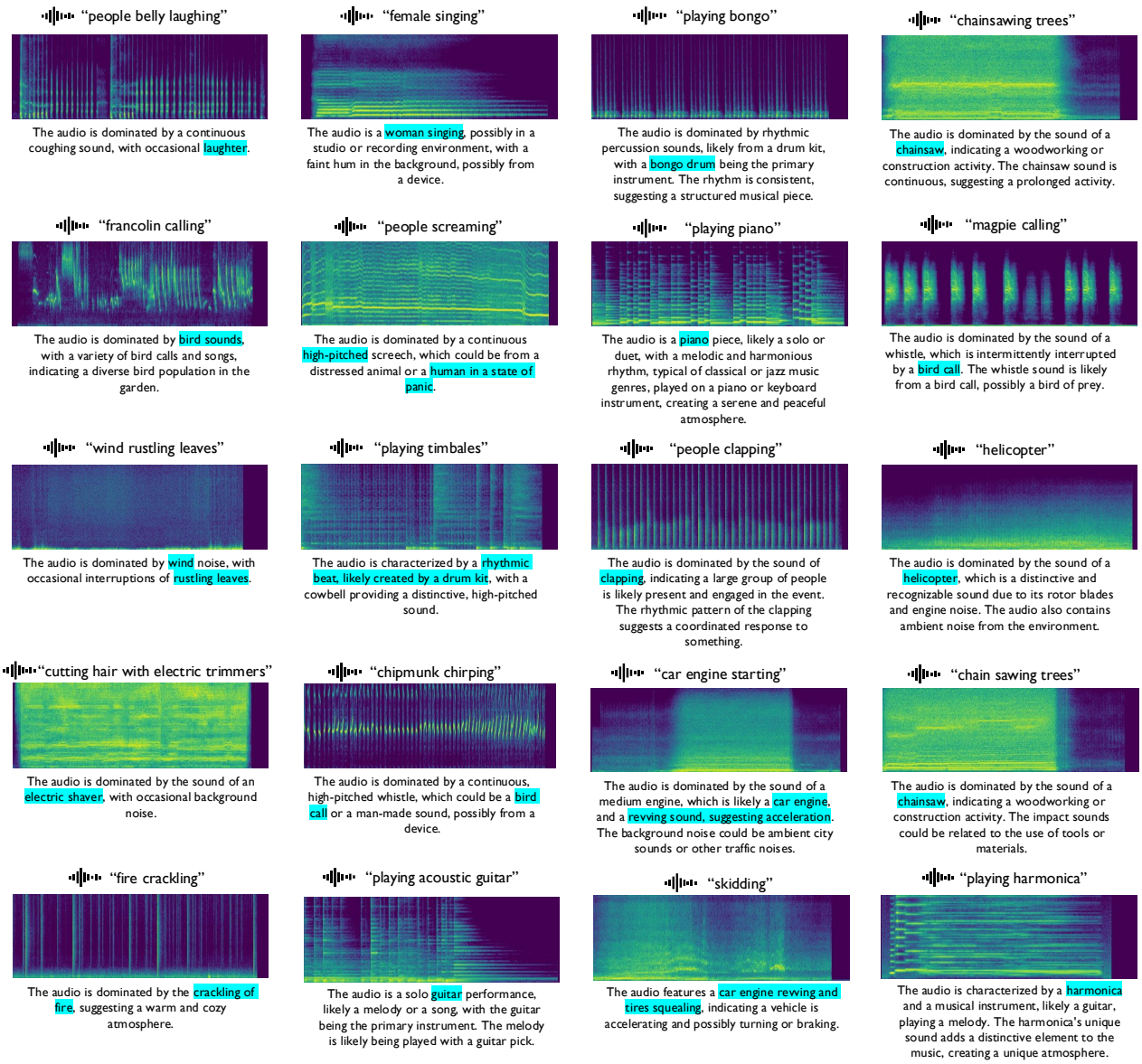

Figure 6. **Generated Audio Examples.**

	Method	cIoU	cIoU Adap.	AUC	AUC Adap.
VGG-SS	Original	44.95 ± 1.14	57.73 ± 1.00	42.18 ± 0.71	49.67 ± 0.51
	Synthetic	43.48 ± 0.79	55.35 ± 0.83	40.80 ± 0.62	47.84 ± 0.45
	(SynI,RealA)	48.32 ± 0.46	59.87 ± 0.94	44.35 ± 0.53	51.00 ± 0.55
	(RealI,SynA)	43.28 ± 4.89	57.20 ± 3.22	41.23 ± 2.51	49.34 ± 2.02
	(SynI,MixedA)	46.35 ± 0.74	57.84 ± 0.80	42.75 ± 0.67	49.51 ± 0.47
	(MixedI,SynA)	45.22 ± 1.38	56.46 ± 1.58	42.36 ± 1.14	48.65 ± 1.02
IS4	Original	57.09 ± 2.58	69.44 ± 2.18	47.36 ± 1.55	56.59 ± 1.43
	Synthetic	58.95 ± 1.20	70.46 ± 1.17	48.78 ± 1.03	57.09 ± 0.74
	(SynI,RealA)	68.71 ± 1.36	79.77 ± 1.24	55.32 ± 0.91	63.13 ± 0.79
	(RealI,SynA)	54.15 ± 3.19	68.45 ± 2.58	46.33 ± 1.80	56.00 ± 1.75
	(SynI,MixedA)	64.40 ± 1.55	75.39 ± 1.81	52.32 ± 0.77	60.28 ± 1.06
	(MixedI,SynA)	61.23 ± 1.82	72.21 ± 1.60	50.83 ± 1.39	58.30 ± 0.93
VPO-SS	Original	39.21 ± 8.58	55.32 ± 3.78	34.32 ± 4.98	46.84 ± 2.21
	Synthetic	39.66 ± 2.58	53.35 ± 2.00	34.46 ± 1.85	45.11 ± 1.29
	(SynI,RealA)	49.27 ± 3.02	60.73 ± 2.97	40.95 ± 1.94	49.93 ± 1.69
	(RealI,SynA)	41.07 ± 7.54	57.47 ± 0.91	36.39 ± 3.26	47.12 ± 0.49
	(SynI,MixedA)	46.32 ± 2.84	59.20 ± 1.91	38.66 ± 1.74	48.58 ± 0.98
	(MixedI,SynA)	44.27 ± 3.21	56.55 ± 1.18	37.59 ± 2.08	47.06 ± 0.85
VPO-MS	Original	39.23 ± 4.78	54.23 ± 1.68	35.19 ± 2.64	46.33 ± 0.92
	Synthetic	36.44 ± 1.81	50.96 ± 1.54	33.59 ± 1.09	44.58 ± 1.04
	(SynI,RealA)	43.58 ± 1.54	56.60 ± 1.13	38.65 ± 1.28	48.21 ± 0.99
	(RealI,SynA)	38.10 ± 4.09	53.64 ± 1.14	35.11 ± 1.88	45.57 ± 0.71
	(SynI,MixedA)	41.41 ± 2.00	55.18 ± 1.48	36.89 ± 1.51	47.01 ± 0.88
	(MixedI,SynA)	40.05 ± 2.38	53.68 ± 1.27	36.21 ± 1.76	46.04 ± 0.96
AVS-S4	Original	59.69 ± 3.59	74.39 ± 2.19	50.71 ± 2.06	60.52 ± 1.36
	Synthetic	61.31 ± 1.84	74.28 ± 1.26	51.67 ± 0.63	60.51 ± 0.77
	(SynI,RealA)	65.77 ± 1.61	78.82 ± 1.60	54.39 ± 1.01	63.02 ± 0.99
	(RealI,SynA)	55.30 ± 2.97	71.56 ± 2.46	48.07 ± 1.71	58.39 ± 1.47
	(SynI,MixedA)	62.70 ± 2.03	76.14 ± 1.28	52.67 ± 1.24	61.47 ± 0.73
	(MixedI,SynA)	62.29 ± 1.76	74.83 ± 1.65	52.83 ± 1.38	60.76 ± 1.17
Avg.	Original	48.03	62.22	41.95	51.99
	Synthetic	47.97	60.88	41.86	51.03
	(SynI,RealA)	55.13	67.16	46.73	55.06
	(RealI,SynA)	46.38	61.66	41.43	51.28
	(SynI,MixedA)	52.24	64.75	44.66	53.37
	(MixedI,SynA)	50.61	62.75	43.96	52.16

	Method	mIoU	mIoU Adap.	F-Score	F-Score Adap.
AVS-S4	Original	55.53 ± 3.08	61.25 ± 2.30	63.65 ± 2.99	68.97 ± 2.30
	Synthetic	54.76 ± 0.50	61.07 ± 0.92	61.79 ± 0.62	69.23 ± 1.07
	(SynI,RealA)	59.81 ± 1.35	65.22 ± 1.26	67.43 ± 1.52	73.49 ± 1.39
	(RealI,SynA)	50.65 ± 2.62	57.49 ± 2.35	58.49 ± 2.50	65.44 ± 2.29
	(SynI,MixedA)	56.47 ± 1.23	62.84 ± 0.97	63.65 ± 1.44	71.06 ± 1.05
	(MixedI,SynA)	56.07 ± 2.23	61.72 ± 1.23	63.19 ± 2.58	69.77 ± 1.45
AVS-MS3	Original	39.59 ± 3.79	42.45 ± 4.10	43.74 ± 4.46	51.54 ± 4.90
	Synthetic	41.44 ± 1.59	42.77 ± 1.52	44.68 ± 1.84	51.27 ± 2.07
	(SynI,RealA)	41.74 ± 1.52	41.91 ± 1.54	45.51 ± 2.13	50.53 ± 1.91
	(RealI,SynA)	39.84 ± 5.25	43.41 ± 4.08	44.61 ± 5.09	52.71 ± 4.11
	(SynI,MixedA)	43.64 ± 1.09	44.60 ± 2.27	47.81 ± 1.29	53.64 ± 2.63
	(MixedI,SynA)	42.81 ± 3.19	43.69 ± 2.58	47.93 ± 3.83	52.34 ± 2.96
Avg.	Original	47.56	51.85	53.70	60.26
	Synthetic	48.10	51.92	53.24	60.25
	(SynI,RealA)	50.78	53.57	56.47	62.01
	(RealI,SynA)	45.25	50.45	51.55	59.08
	(SynI,MixedA)	50.06	53.72	55.73	62.35
	(MixedI,SynA)	49.44	52.71	55.56	61.06

	Method	IIoU	IIoU Adap.	IAUC	IAUC Adap.
IS4	Original	31.78 ± 2.78	45.14 ± 4.92	30.53 ± 1.69	40.74 ± 1.91
	Synthetic	34.64 ± 1.35	49.66 ± 1.66	32.76 ± 1.30	42.01 ± 1.01
	(SynI,RealA)	46.89 ± 1.99	63.54 ± 1.98	41.27 ± 1.35	50.63 ± 1.25
	(RealI,SynA)	29.15 ± 3.06	46.64 ± 3.67	29.85 ± 2.06	40.29 ± 2.39
	(SynI,MixedA)	41.00 ± 2.30	56.60 ± 2.92	37.07 ± 1.13	46.34 ± 1.71
	(MixedI,SynA)	36.85 ± 1.79	52.07 ± 2.11	35.23 ± 1.87	43.52 ± 1.39
VPO-MS	Original	30.41 ± 6.15	46.27 ± 2.27	28.40 ± 3.16	40.05 ± 1.27
	Synthetic	28.11 ± 1.78	43.22 ± 1.51	27.13 ± 1.44	38.46 ± 1.04
	(SynI,RealA)	36.09 ± 2.00	49.68 ± 1.39	32.74 ± 1.51	42.74 ± 1.12
	(RealI,SynA)	30.96 ± 5.21	47.16 ± 1.24	29.91 ± 1.95	40.28 ± 0.68
	(SynI,MixedA)	34.29 ± 2.70	48.61 ± 1.87	31.17 ± 2.04	41.55 ± 1.06
	(MixedI,SynA)	32.59 ± 2.87	46.74 ± 1.25	30.07 ± 1.87	40.34 ± 0.84
Avg.	Original	31.10	45.71	29.47	40.40
	Synthetic	31.38	46.44	29.95	40.24
	(SynI,RealA)	41.49	56.61	37.01	46.69
	(RealI,SynA)	30.06	46.90	29.88	40.29
	(SynI,MixedA)	37.65	52.61	34.12	43.95
	(MixedI,SynA)	34.72	49.41	32.65	41.93

Table 5. **Sound source localization results on the same data scale.** *Mixed* denotes both synthetic and real samples within the same modality are used in a mixed form as (134K Syn. + 10K Real).

	Method	cIoU	cIoU Adap.	AUC	AUC Adap.
VGG-SS	(A) {(S, S) ∪ (S, S)}	43.20 ± 0.76	54.86 ± 1.06	40.75 ± 0.38	47.65 ± 0.78
	(B) {(S, R) ∪ (S, R)}	48.02 ± 0.99	59.55 ± 1.01	44.16 ± 0.77	50.86 ± 0.80
	(C) {(R, R) ∪ (S, R)}	52.19 ± 1.82	63.37 ± 1.89	46.57 ± 0.98	53.07 ± 1.06
	(D) {(R, R) ∪ (S, S)}	48.92 ± 0.54	61.38 ± 1.02	45.21 ± 0.49	52.41 ± 0.74
	(E) {(S, R) ∪ (S, S)}	48.39 ± 0.71	60.31 ± 0.73	44.13 ± 0.69	51.16 ± 0.37
IS4	(A) {(S, S) ∪ (S, S)}	58.63 ± 1.55	70.23 ± 1.47	48.83 ± 0.62	57.01 ± 0.97
	(B) {(S, R) ∪ (S, R)}	69.02 ± 0.83	80.08 ± 1.31	55.33 ± 0.63	63.26 ± 0.75
	(C) {(R, R) ∪ (S, R)}	68.44 ± 1.11	79.90 ± 0.71	54.75 ± 0.56	63.35 ± 0.47
	(D) {(R, R) ∪ (S, S)}	64.28 ± 0.60	76.29 ± 0.68	52.27 ± 0.40	60.98 ± 0.49
	(E) {(S, R) ∪ (S, S)}	69.24 ± 0.80	80.79 ± 0.79	55.40 ± 0.45	63.63 ± 0.53
VPO-SS	(A) {(S, S) ∪ (S, S)}	39.08 ± 1.78	52.47 ± 2.56	34.48 ± 0.90	44.50 ± 1.89
	(B) {(S, R) ∪ (S, R)}	47.50 ± 3.45	60.47 ± 2.63	40.32 ± 2.50	49.69 ± 1.39
	(C) {(R, R) ∪ (S, R)}	48.67 ± 4.28	62.08 ± 2.57	40.94 ± 2.44	51.43 ± 1.54
	(D) {(R, R) ∪ (S, S)}	42.32 ± 3.57	56.03 ± 2.72	36.46 ± 2.49	47.65 ± 1.45
	(E) {(S, R) ∪ (S, S)}	48.09 ± 2.85	61.46 ± 1.18	40.66 ± 1.41	50.69 ± 0.77
VPO-MS	(A) {(S, S) ∪ (S, S)}	35.88 ± 0.89	49.89 ± 1.56	33.34 ± 0.51	43.74 ± 1.07
	(B) {(S, R) ∪ (S, R)}	43.37 ± 1.95	56.47 ± 1.43	38.34 ± 1.38	48.24 ± 0.80
	(C) {(R, R) ∪ (S, R)}	44.30 ± 1.98	57.85 ± 1.69	39.43 ± 1.31	49.18 ± 1.00
	(D) {(R, R) ∪ (S, S)}	40.35 ± 2.11	54.31 ± 1.60	36.24 ± 1.21	46.80 ± 0.95
	(E) {(S, R) ∪ (S, S)}	43.59 ± 1.21	57.27 ± 0.41	38.68 ± 0.48	48.78 ± 0.32
AVS-S4	(A) {(S, S) ∪ (S, S)}	58.77 ± 2.14	72.82 ± 2.05	50.24 ± 1.56	59.63 ± 1.83
	(B) {(S, R) ∪ (S, R)}	65.51 ± 2.18	78.24 ± 1.47	54.18 ± 1.24	62.84 ± 0.81
	(C) {(R, R) ∪ (S, R)}	68.19 ± 1.30	81.14 ± 1.35	55.76 ± 0.75	64.55 ± 1.00
	(D) {(R, R) ∪ (S, S)}	66.77 ± 1.23	80.06 ± 1.05	55.28 ± 0.86	64.21 ± 1.01
	(E) {(S, R) ∪ (S, S)}	66.55 ± 1.12	79.29 ± 0.89	54.75 ± 0.64	63.67 ± 0.58
Avg.	(A) {(S, S) ∪ (S, S)}	47.11	60.05	41.53	50.51
	(B) {(S, R) ∪ (S, R)}	54.68	66.96	46.47	54.98
	(C) {(R, R) ∪ (S, R)}	56.36	68.87	47.49	56.32
	(D) {(R, R) ∪ (S, S)}	52.53	65.61	45.09	54.41
	(E) {(S, R) ∪ (S, S)}	55.17	67.82	46.72	55.59

	Method	mIoU	mIoU Adap.	F-Score	F-Score Adap.
AVS-S4	(A) {(S, S) ∪ (S, S)}	52.84 ± 2.46	59.88 ± 2.35	59.86 ± 2.87	68.13 ± 2.63
	(B) {(S, R) ∪ (S, R)}	58.50 ± 1.57	64.67 ± 0.99	66.06 ± 1.76	73.14 ± 1.13
	(C) {(R, R) ∪ (S, R)}	61.71 ± 0.92	66.38 ± 1.33	69.86 ± 0.98	74.63 ± 1.52
	(D) {(R, R) ∪ (S, S)}	60.55 ± 1.09	65.80 ± 1.08	68.57 ± 1.01	74.09 ± 1.27
	(E) {(S, R) ∪ (S, S)}	59.28 ± 1.40	65.50 ± 0.79	67.07 ± 1.43	74.00 ± 0.89
AVS-MS3	(A) {(S, S) ∪ (S, S)}	38.35 ± 1.63	39.85 ± 2.22	40.48 ± 2.00	47.78 ± 2.73
	(B) {(S, R) ∪ (S, R)}	42.12 ± 2.96	43.23 ± 2.55	45.74 ± 3.81	52.21 ± 3.25
	(C) {(R, R) ∪ (S, R)}	44.05 ± 2.50	45.27 ± 2.81	49.15 ± 2.34	54.98 ± 2.90
	(D) {(R, R) ∪ (S, S)}	45.07 ± 1.86	46.77 ± 1.85	49.83 ± 2.37	56.28 ± 2.27
	(E) {(S, R) ∪ (S, S)}	43.59 ± 2.34	45.01 ± 1.86	47.87 ± 3.38	54.75 ± 2.42
Avg.	(A) {(S, S) ∪ (S, S)}	45.60	49.87	50.17	57.96
	(B) {(S, R) ∪ (S, R)}	50.31	53.95	55.90	62.68
	(C) {(R, R) ∪ (S, R)}	52.88	55.83	59.51	64.81
	(D) {(R, R) ∪ (S, S)}	52.81	56.29	59.20	65.19
	(E) {(S, R) ∪ (S, S)}	51.44	55.26	57.47	64.38

	Method	IIoU	IIoU Adap.	IAUC	IAUC Adap.
IS4	(A) {(S, S) ∪ (S, S)}	34.25 ± 1.82	49.15 ± 1.81	32.66 ± 0.86	41.88 ± 1.14
	(B) {(S, R) ∪ (S, R)}	47.48 ± 1.66	64.05 ± 2.24	41.27 ± 1.09	50.86 ± 1.30
	(C) {(R, R) ∪ (S, R)}	46.22 ± 1.50	63.75 ± 1.19	40.19 ± 0.69	50.74 ± 0.78
	(D) {(R, R) ∪ (S, S)}	40.89 ± 0.60	58.07 ± 0.95	36.97 ± 0.59	47.36 ± 0.69
	(E) {(S, R) ∪ (S, S)}	48.03 ± 1.23	65.09 ± 1.41	41.48 ± 0.61	51.42 ± 0.93
VPO-MS	(A) {(S, S) ∪ (S, S)}	27.82 ± 1.19	42.46 ± 2.05	27.16 ± 0.47	37.89 ± 1.28
	(B) {(S, R) ∪ (S, R)}	35.79 ± 2.49	49.61 ± 1.63	32.37 ± 1.82	42.73 ± 0.86
	(C) {(R, R) ∪ (S, R)}	36.47 ± 2.66	50.82 ± 1.92	33.31 ± 1.70	43.56 ± 1.18
	(D) {(R, R) ∪ (S, S)}	31.54 ± 2.57	46.42 ± 2.14	29.65 ± 1.42	40.71 ± 1.21
	(E) {(S, R) ∪ (S, S)}	36.03 ± 1.76	50.44 ± 0.59	32.78 ± 0.80	43.37 ± 0.36
Avg.	(A) {(S, S) ∪ (S, S)}	31.04	45.81	29.91	39.89
	(B) {(S, R) ∪ (S, R)}	41.64	56.83	36.82	46.80
	(C) {(R, R) ∪ (S, R)}	41.35	57.29	36.75	47.15
	(D) {(R, R) ∪ (S, S)}	36.22	52.25	33.31	44.04
	(E) {(S, R) ∪ (S, S)}	42.03	57.77	37.13	47.40

Table 6. **Sound source localization results on 2x scaled data.** R: 144K Real Images, S: 144K Synthetic Images, S: 144K Synthetic Images, R: 144K Real Audios, S: 144K Synthetic Audios, S: 144K Synthetic Audios.

	Method	cIoU	cIoU Adap.	AUC	AUC Adap.
VGG-SS	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	43.19 ± 1.02	56.03 ± 0.94	40.72 ± 1.43	48.40 ± 0.61
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	48.89 ± 0.97	60.52 ± 1.31	44.64 ± 0.74	51.12 ± 0.84
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	51.37 ± 0.57	62.51 ± 0.70	46.16 ± 0.39	52.47 ± 0.52
IS4	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	59.63 ± 1.99	71.93 ± 1.46	49.10 ± 2.13	58.01 ± 0.84
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	71.02 ± 2.19	82.06 ± 2.18	56.38 ± 1.08	64.42 ± 1.09
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	71.54 ± 1.16	82.27 ± 0.82	56.55 ± 0.56	64.65 ± 0.50
VPO-SS	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	42.03 ± 4.60	54.01 ± 3.68	35.99 ± 2.85	45.81 ± 2.38
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	48.85 ± 4.07	62.33 ± 1.91	41.23 ± 2.25	51.11 ± 1.05
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	48.43 ± 2.34	62.40 ± 1.88	41.03 ± 1.31	51.58 ± 1.22
VPOMS	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	37.74 ± 4.19	50.88 ± 3.59	34.21 ± 3.19	44.50 ± 2.31
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	43.54 ± 3.41	56.95 ± 1.36	38.89 ± 1.99	48.80 ± 0.75
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	43.84 ± 2.12	57.84 ± 2.57	38.84 ± 1.51	49.35 ± 1.42
AVS-S4	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	61.94 ± 1.89	75.57 ± 1.25	52.04 ± 1.57	61.60 ± 1.02
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	66.33 ± 0.99	79.21 ± 0.82	54.61 ± 0.39	63.39 ± 0.40
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	68.11 ± 1.71	81.11 ± 1.39	55.73 ± 0.94	64.63 ± 0.86
Avg.	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	48.91	61.68	42.41	51.66
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	55.73	68.21	47.15	55.77
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	56.66	69.23	47.66	56.54

	Method	mIoU	mIoU Adap.	F-Score	F-Score Adap.
AVS-S4	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	55.31 ± 2.27	62.80 ± 1.18	62.70 ± 2.42	71.21 ± 1.19
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	58.88 ± 0.39	64.88 ± 0.47	66.65 ± 0.43	73.29 ± 0.34
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	61.66 ± 1.19	66.54 ± 0.85	69.74 ± 1.28	74.72 ± 0.92
AVSMS3	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	43.24 ± 1.05	44.97 ± 1.65	46.93 ± 1.14	53.60 ± 2.54
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	44.27 ± 2.01	45.21 ± 1.52	48.88 ± 2.33	54.50 ± 1.49
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	44.61 ± 2.03	45.59 ± 1.99	49.20 ± 2.38	55.22 ± 2.52
Avg.	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	49.28	53.89	54.82	62.41
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	51.58	55.05	57.77	63.90
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	53.14	56.07	59.47	64.97

	Method	IIoU	IIoU Adap.	IAUC	IAUC Adap.
IS4	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	35.40 ± 2.07	51.73 ± 1.48	33.32 ± 2.51	43.41 ± 0.79
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	50.29 ± 3.33	67.29 ± 3.58	42.68 ± 1.70	52.66 ± 1.79
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	50.78 ± 1.68	67.51 ± 1.39	42.74 ± 0.67	52.84 ± 0.76
VPOMS	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	29.78 ± 4.92	43.21 ± 3.79	28.25 ± 3.37	38.58 ± 2.41
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	36.00 ± 4.49	49.99 ± 2.02	33.05 ± 2.44	43.29 ± 0.91
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	35.73 ± 2.27	50.59 ± 2.65	32.67 ± 1.54	43.74 ± 1.46
Avg.	(A) {(S, S) ∪ (S, S) ∪ (S, S)}	32.59	47.47	30.79	41.00
	(B) {(S, R) ∪ (S, R) ∪ (S, R)}	43.15	58.64	37.87	47.98
	(C) {(R, R) ∪ (S, R) ∪ (S, R)}	43.26	59.05	37.71	48.29

Table 7. **Sound source localization results on 3× scaled data.** `VGGSound` (R: 144K Real Images, R: 144K Real Audios), `VGGSyn1` (S: 144K Synthetic Images, S: 144K Synthetic Audios), `VGGSyn2` (S: 144K Synthetic Images, S: 144K Synthetic Audios) and `VGGSyn3` (S: 144K Synthetic Images, S: 144K Synthetic Audios).

References

[1] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, 2021. 1

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 3

[3] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP*, 2025. 3, 4

[4] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *EMNLP*, 2024. 5

[5] 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*, 2023. 3

[6] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3

[7] Jia Li, Wenjie Zhao, Ziru Huang, Yunhui Guo, and Yapeng Tian. Do audio-visual segmentation models truly segment sounding objects? *arXiv preprint arXiv:2502.00358*, 2025. 1, 2, 3

[8] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *ACM MM*, 2024. 3, 4

[9] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *NeurIPS*, 2022. 3

[10] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 2

[11] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *WACV*, 2022. 3

[12] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Toward interactive sound source localization: Better align sight and sound! *IEEE TPAMI*, 2025. 1, 2

[13] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *CVPR*, 2023. 3

[14] Chen Yuanhong, Liu Yuyuan, Wang Hu, Liu Fengbei, Wang Chong, and Carneiro Gustavo. Unraveling instance associations: A closer look for audio-visual segmentation. In *CVPR*, 2024. 1

[15] 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 *ECCV*, 2022. 1, 2