SAVGBench: Benchmarking Spatially Aligned Audio-Video Generation

Kazuki Shimada[1,*], Christian Simon[2,*], Takashi Shibuya[1], Shusuke Takahashi[2], Yuki Mitsufuji[1,2]
[1] Sony AI, [2] Sony Group Corporation

Abstract

This work addresses the lack of generative models that produce high-quality videos with spatially aligned audio, a key aspect often overlooked in recent advancements. To tackle this problem, we establish a new research direction in benchmarking Spatially Aligned Audio-Video Generation (SAVG). We propose three key components for the benchmark: dataset, baseline, and metrics. We introduce a spatially aligned audio-visual dataset, derived from an audio-visual dataset consisting of multichannel audio, video, and spatiotemporal annotations of sound events. We propose a baseline audio-visual diffusion model focused on stereo audio-visual joint learning to accommodate spatial sound. Finally, we present metrics to evaluate video and spatial audio quality, including a new spatial audio-visual alignment metric. Our experimental result demonstrates that gaps exist between the baseline model and ground truth in terms of video and audio quality, and spatial alignment between both modalities.

1. Introduction

Recently, generative models (*e.g*., diffusion models and transformer-based models) have shown remarkable achievement in generating high-quality videos [8, 10, 5]. However, there are only a few models targeting multimodal generation, especially samples with audio-visual elements [8, 13]. Furthermore, videos generated by current state-of-the-art (SOTA) techniques, *e.g*., [10, 5, 8], fail to accurately represent real-world conditions, as they overlook spatial information necessary for creating immersive content. Consequently, the generated videos often lack the realism required to spatially align audio with visual elements. The spatial component of a video not only enhances realism for experiential purposes but also provides a contextual understanding of the video (*i.e*., sound source directions), which can be used for various applications, *e.g*., virtual reality, world simulation, and robot perception.

This work establishes a new research direction in benchmarking Spatially Aligned Audio-Video Generation (SAVG), illustrated in Fig. 1. We begin with stereo audio and perspective video formats, widely used in media content. This area remains underexplored, with a lack of data and standardized benchmarks. To advance this direction, we introduce SAVGBench, a benchmark with three core components: dataset, baseline, and metrics.

* Equal contribution

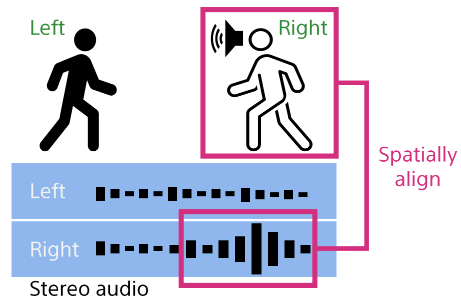

Figure 1. An illustration of a spatially aligned audio-video generation (SAVG).

The essence for generating spatially aligned audio and video is the training data, either used for training from scratch or finetuning. In this work, we extend a dataset called Sony-TAU Realistic Spatial Soundscapes 2023 (STARSS23) [9], which consists of spatially aligned Ambisonics audio and 360° video data with oracle position labels of sound events. This Ambisonics audio data, 360° video data, and position labels provide omnidirectional coverage around the microphone array and camera, allowing us to track the position of sound events on and off the screen when we convert them into stereo audio and perspective video. We create an audio-visual dataset containing stereo audio and perspective video data, curated based on onscreen and offscreen events.

The second is a baseline model. A key to building a model with spatial alignment is learning a joint distribution over both modalities. The design of MM-Diffusion [8] enables such audio-visual joint learning. We propose a stereo extension of the diffusion model to achieve joint learning between stereo audio and video.

The third is how to evaluate generated audio and video. We use the Fréchet video distance (FVD) [11] and Fréchet audio distance (FAD) [4] metrics to assess video and audio quality. To evaluate spatial alignment between video and audio, we introduce a new metric based on detecting sounding object positions in both modalities and measuring their alignment. The metric relies on object detection [2] and sound event localization and detection (SELD) [12].

In this work, our contributions are three-fold:

1. We introduce an SAVG dataset containing videos with a perspective view and stereo audio, converted and curated from STARSS23.
2. We also propose a stereo audio-visual diffusion model designed to address audio-visual joint learning with a focus on spatial sound.

Figure 2. Examples of our proposed dataset with various indoor environments and sound events.

3. We also introduce a set of metrics to assess the quality of videos and spatial sound generated from a model. In particular, we propose a spatial audio-visual alignment metric using an object detector and a SELD model.

2. Proposed Dataset: SVGSA24

2.1. Overview

We create an audio-visual dataset named **SVGSA24** that contains stereo audio and perspective video data, derived from the STARSS23 dataset [9]. STARSS23 consists of sound scene recordings with various rooms and sound events, containing first-order Ambisonics (FOA) audio data, the corresponding equirectangular video data, and spatiotemporal annotations, *i.e.*, classes, activities, and positions of sound events. We convert the STARSS23 data into stereo audio and perspective video data, tracking the position of sound events on the screen. We provide examples of the proposed dataset in Fig. 2.

The SVGSA24 dataset features humans and musical instruments in indoor environments, including speeches and instrument sounds. The audio data are delivered as stereo audio with a 16 kHz sampling rate. The video data uses a perspective view, ensuring that a video reflects sound events in the audio. The video resolution is 256×256 with padding. The length and frame per second (fps) rate are set to 5 seconds and 4 fps, respectively.

SVGSA24 is split into the development and evaluation sets. The development set contains 5,031 videos, totaling about 7 hours. We release the development set to the public[1] and keep the evaluation set for a challenge evaluation[2]. The evaluation set serves as a target distribution to quantify the quality of generated video and audio. In all sets, the ratio of speech to instrument sounds is maintained at 2:1.

2.2. Data Construction

The SVGSA24 dataset is constructed as follows. First, we extract 5-second videos every 0.5 seconds from the STARSS23 data. Then, we convert the videos with the equirectangular view and FOA audio to videos with a perspective view and stereo audio, using a fixed viewing angle for the perspective view. While we keep the vertical viewing angle at 0 degrees, we change the horizontal viewing angle by 10 degrees, sampling videos with a perspective view and stereo audio. Finally, we curate videos that contain only onscreen speech and instrument sounds. Note that the step sizes (*i.e.*, 0.5 seconds and 10 degrees) are for the development set. We use different step sizes for the evaluation set to keep the ratio of speech to instrument sounds at the same ratio as the development set.

Data Conversion. Using a fixed viewing angle, we convert the equirectangular view and FOA audio to a perspective view and stereo audio. **FOA → stereo:** According to the viewing angle, we first rotate the FOA audio [7]. Then we convert the rotated FOA audio to the stereo with a simple translation [12]: $\texttt{left} = W+Y$ and $\texttt{right} = W-Y$, where W is the omnidirectional signal of the FOA audio, and Y is the first-order horizontal (left-right) component [12]. **Equirectangular → perspective:** We convert the equirectangular video to a perspective video with the same viewing angle as the audio, using a python library[3]. We set the horizontal field of view to 100 degrees. We also set the output height and width to 144 and 256, whose aspect ratio is 16:9. After the conversion, we add padding to make a video with 256×256 resolution, to follow the pretrained super-resolution model's resolution settings.

Data Curation. We curate videos that contain only onscreen sound events. Including only onscreen events facilitates the evaluation of SAVG. During data conversion, we also convert position labels in the equirectangular video to position labels in a new perspective video. Using the position labels in the perspective video, we investigate whether an event is onscreen or offscreen.

We also curate only events from speech and instrument classes, although the STASS23 dataset contains other sound event classes. The speech and instrument classes are stably detected by the object detector and SELD model, leading to a stable evaluation of spatial alignment. In our preliminary experiments with the object detector, the person class was well detected, while other classes, such as cell phone or sink, were not reliably detected in the 256×256 resolution videos. So, we considered using only human body-related classes, *i.e.*, speech, clap, laugh, footstep, and instrument (as humans play instruments). On the other hand, when we trained a SELD model with the human body-related classes, the SELD model did not perform well in detecting the clap, laugh, and footstep classes. Finally, we use the speech and instrument classes to create our proposed dataset. We use class labels for each event to investigate whether an event belongs to these target classes.

Other Procedures. In addition to data conversion and curation, several steps are performed to produce the final version of the proposed dataset. We remove videos with over-

[1] https://drive.google.com/file/d/14Fy6C_N6BXymYKhXMxVbt7tHnZmVRMEd/view

[2] https://www.aicrowd.com/challenges/sounding-video-generation-svg-challenge-2024/problems/spatial-alignment-track

[3] https://github.com/sunset1995/py360convert

lapping events to focus on single source cases. We set a threshold for the total length of sound events in a video to 80% to ensure that the video contains sufficient sound events. When the total length of sound events is less than 4 seconds in a 5-second video, we remove the video. We apply a high-pass filter to all the audio data and amplify it by 38 dB to enhance its sound event signals, stabilizing the training of our baseline model. If the amplified audio of a video is clipping, we remove the video.

3. Baseline Model: Stereo MM-Diffusion

A key to building a model with spatial alignment for diffusion models is learning a joint distribution over audio and visual modalities. In this work, we propose a stereo channel extension of MM-Diffusion [8]. We use a structure similar to that of MM-Diffusion. The learning mechanism in MM-Diffusion implements joint learning between audio and video. Thus, the gaps between generated outputs between the two modalities are narrowed down.

In our implementation, MM-Diffusion consists of two separate branches for audio and video processing. In particular, the model encodes the source audio waveform using an audio encoder, yielding stereo audio with the size of $2 \times C \times T$, where C and T are the feature channel and corresponding time sequence. In the video branch, an encoder maps a video sequence of F frames to dimensions of $F \times C \times H \times W$, where H and W are height and width of a frame. The outputs from both the audio and video encoders are then integrated through a multi-modal attention module. This approach enables better alignment between audio and video, compared to training each modality independently. We follow the architectural setup of the original MM-Diffusion [8], using 4 scales of MM-Blocks, each comprising 2 standard MM-Blocks, along with an additional downsample or upsample block.

As the model requires a huge amount of GPU memory, we only deal with videos of size 64×64 to fit a sample in a single GPU. In audio-visual generation with spatial alignment, precise positioning of objects is essential. An approach to obtaining object positions is to use an object detector. However, at the 64×64 resolution, the sounding objects are not visible, either to the object detector or even to the human eye. Therefore, a super-resolution model is required to upsample the video to 256×256 for better visibility of objects. The super-resolution model uses an identical architecture as in the guided diffusion model [1].

As the model is designed unconditionally, the trained model could be used to generate a pair of audio-video samples. We observe that the DDPM [3], along with the MM-Diffusion model, is slow to generate a sample. To expedite the testing process, we use DPM Solver [6]. There is a trade-off in quality, but this issue is not significant as we only generate for a small resolution 64×64. For the super-

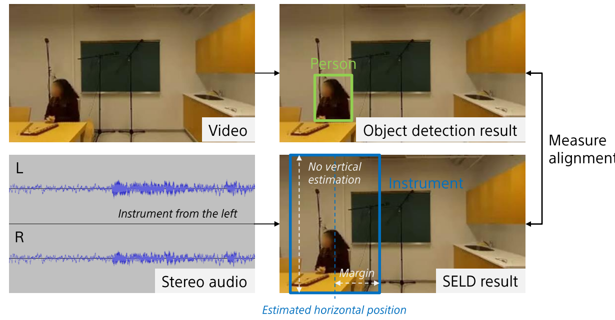

Figure 3. An illustration of Spatial AV-Align metric with the detected object (person class) and the detected sound event (instrument class). The green box indicates the detected object using the object detector. The blue box indicates the detected sound event using the SELD model. The SELD result has a margin around the estimated horizontal position. Its vertical range is set from the top to the bottom as it does not estimate a vertical position.

resolution model, we still use DDPM to maintain the quality of the generated video.

4. Evaluation Metrics

We introduce a set of metrics to assess the quality of video and spatial audio generated from a model. To evaluate the quality of audio-visual samples, we use the FVD [11] and FAD [4] metrics to assess video and audio quality, respectively. To measure spatial alignment between an audio sample and the corresponding video sample, we propose a new metric, **Spatial AV-Align**, which ranges from zero to one, with higher values indicating better alignment.

The Spatial AV-Align metric evaluates spatial synchronization between sound events in audio and objects in video. The metric relies on prior works on object detection [2] and sound event localization and detection (SELD) [12]. We use the widely known object detector YOLOX [2] to identify objects in video. To detect and localize sound events in audio, we prepare a stereo SELD model [12], which takes stereo audio as an input to estimate each class's activity and horizontal position within the video per frame. The model is trained with binary cross entropy and mean squared error using the development set of the SVGSA24 dataset. It is evaluated using the evaluation set, achieving over 0.95 in each class's F-score. Please see Fig. 3 for an illustration of a detected object (person class) and a detected sound event (instrument class) from each input. Note that the SELD result has a fixed margin around the estimated value in a horizontal position. The vertical range of a SELD result is set from the top to the bottom as it does not have a vertical estimation. To evaluate this metric using stably detected objects or sound events, we focus on the person class in object detection and speech and instrument classes in SELD, as in the dataset section.

We explain the specific flow to compute the metric. We begin by detecting candidate positions of sounding objects

Model	FVD ↓	FAD ↓	Spatial AV-Align ↑
MM-Diffusion	1050.3	9.65	0.48
Ground Truth	572.05	3.70	0.92

Table 1. Evaluation results of the baseline and ground truth on key metrics. FVD and FAD evaluate the quality of video and audio, respectively, while our proposed Spatial AV Align metric indicates the alignment between both modalities.

per frame within each modality separately. Note that the fps for each modality is different: the object detector outputs at 4 fps, whereas the SELD model outputs at 10 fps. Afterward, for a detected position in an audio frame, we verify if an object is also detected at the same position in the closest video frame to the audio frame. Specifically, we determine if the SELD result overlaps an object detection result. If there is an overlap, it is considered as a true positive; otherwise, it is a false negative. We allow using object detection results across the adjacent video frames (*i.e.*, from the previous to the next) to account for temporal context. We do not verify whether a detected object in a video frame appears in the corresponding audio frame since the dataset includes people who do not speak or play instruments. Finally, we compute a recall metric as the alignment score, ranging from zero to one. This alignment score is defined as $TP/(TP+FN)$, where TP and FN indicate numbers of true positives and false negatives, respectively.

5. Experimental Evaluation

In our experiments, we use the development set of the SVGSA24 dataset to train the Stereo MM-Diffusion model with its super-resolution model. The Stereo MM-Diffusion is trained from scratch while the super-resolution model is initialized with a pretrained model from the guided diffusion model [1], which is pretrained on ImageNet. To finetune the super-resolution model, we extract all videos as frames and train the model on each frame. We set a batch size to 4 with 8 NVIDIA A100 GPUs to train both models. In evaluation, we use the evaluation set of SVGSA24 as a reference to measure the quality of generated models.

We evaluate the quality and spatial alignment as shown in Table 1. The MM-Diffusion baseline achieves the Spatial AV-Align metric 0.48. Compared to the Spatial AV-Align score on the evaluation set (*i.e.*, Ground-Truth), this result indicates potential for further improvement.

6. Conclusion

This paper presents a new benchmark for Spatially Aligned Audio-Video Generation (SAVG). We propose three key components for the benchmark: dataset, baseline, and metrics. We introduce a spatially aligned audio-visual dataset named SVGSA24, which enables us to train and evaluate a model that generates videos with spatially aligned stereo audio. We also propose an audio-visual diffusion model focused on stereo audio-visual learning to accommodate spatial sound. We also introduce a new metric, Spatial AV-Align, which evaluates spatial alignment between audio and video using an object detector and a sound event localization and detection (SELD) model. Our experimental result shows that gaps exist between the baseline model and ground truth regarding video and audio quality, and spatial alignment between both modalities. This benchmark encourages future work in SAVG.

References

[1] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Proc. of NeurIPS*, 2021. 3, 4

[2] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv:2107.08430*, 2021. 1, 3

[3] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020. 3

[4] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv:1812.08466*, 2018. 1, 3

[5] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv:2402.17177*, 2024. 1

[6] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Proc. of NeurIPS*, 2022. 3

[7] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada. First order Ambisonics domain spatial augumentation for DNN-based direction of arrival estimation. In *Proc. of DCASE Workshop*, 2019. 2

[8] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo. MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proc. of CVPR*, 2023. 1, 3

[9] K. Shimada, A. Politis, P. Sudarsanam, D. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, et al. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Proc. of NeurIPS*, 2023. 1, 2

[10] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-A-Video: Text-to-video generation without text-video data. *arXiv:2209.14792*, 2022. 1

[11] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. FVD: A new metric for video generation. In *Proc. of ICLR Workshop*, 2019. 1, 3

[12] J. Wilkins, M. Fuentes, L. Bondi, S. Ghaffarzadegan, A. Abavisani, and J. P. Bello. Two vs. four-channel sound event localization and detection. In *Proc. of DCASE Workshop*, 2023. 1, 2, 3

[13] Y. Xing, Y. He, Z. Tian, X. Wang, and Q. Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proc. of CVPR*, 2024. 1