SEE-2-SOUND 🔊: Zero-Shot Spatial Environment-to-Spatial Sound

Rishit Dagli [1] **Shivesh Prakash** [1] **Robert Wu** [1] **Houman Khosravani** [1 2 3]

see2sound.github.io

Abstract

Generating combined visual and auditory sensory experiences is critical for the consumption of immersive content. Recent advances in neural generative models have enabled the creation of high-resolution content across multiple modalities such as images, text, speech, and videos. Despite these successes, there remains a significant gap in the generation of high-quality spatial audio that complements generated visual content. Furthermore, current audio generation models excel in either generating natural audio or speech or music but fall short in integrating spatial audio cues necessary for immersive experiences. In this work, we introduce SEE-2-SOUND, a zero-shot approach that decomposes the task into (1) identifying visual regions of interest; (2) locating these elements in 3D space; (3) generating mono-audio for each; and (4) integrating them into spatial audio. Using our framework, we demonstrate compelling results for generating spatial audio for high-quality videos, images, and dynamic images from the Internet, as well as media generated by learned approaches.

1. Introduction

Human visual and auditory systems inherently depend on synchronized spatial audio cues for an immersive experience and contextual understanding of video content. Modern generative models have been able to create compelling high-resolution digital content across various modalities. These models have been successful in generating images (Saharia et al., 2022; Rombach et al., 2022), text (Raffel et al., 2020; Achiam et al., 2023), speech (van den Oord et al., 2016; Ren et al., 2021), audio (Liu et al., 2024; Kreuk et al., 2023), 3D scenes (Lin et al., 2023; Poole et al., 2023), 4D (dynamic 3D) scenes (Singer et al., 2023b; Ling et al., 2023), and videos (Singer et al., 2023a; Brooks et al., 2024). Although generating audio from text (Huang et al., 2023b; Copet et al., 2024) - and more recently, images (Tang et al., 2024) - has become a reality, generating spatial audio is still a challenge. Unlike earlier works, we are interested in solving the specific problem of generating high-quality spatial audio or surround sound to accompany images, animated images (GIFs), and videos. Such a model could be particularly useful to pair with any kind of video generation or image generation techniques as well as on real images and videos.

Audio generation is a long-standing challenge, often subdivided into generating speech, music, and sound effects. Most audio generation methods include inductive biases for each of these tasks, often due to the diversity between the generation of speech, music, and other types of audio (Bresin et al., 2010; Agostinelli et al., 2023; Liu et al., 2023; Tan et al., 2024). There have also been newer approaches that develop unified representations (Liu et al., 2024) to generate complex audio as well. However, none of these currently emphasize the generation of rich spatial audio. The few related works that perform image-to-audio (Tang et al., 2024; 2023) are closest to our work. However, these also face the same problem of not being able to generate rich spatial audio. Currently, there also exist a few learned approaches that generate spatial audio (Morgado et al., 2018; Lin & Wang, 2021) but these are often based on having access to the captured audio; following which these approaches localized the audio in a viewing sphere to produce spatial audio for a captured video. However, none of these approaches aim to generate accompanying spatial audio for images, animated images (GIFs), and videos, which remains an open problem.

Towards this goal, we propose SEE-2-SOUND, a training-free method that approaches this specific problem by employing the foundational models, CoDi (Tang et al., 2024), DepthAnything (Yang et al., 2024), and Segment Anything (Kirillov et al., 2023). The pipeline works by explicitly factorizing this generative problem into separate problems: (1) identifying regions of interest in the input; (2) finding

[1]University of Toronto, Toronto, Canada [2]Temerty Centre for Artificial Intelligence Research and Education in Medicine [3]Sunnybrook Research Institute, Hurvitz Brain Sciences Program, Division of Neurology, Department of Medicine, University of Toronto. Correspondence to: Rishit Dagli <rishit@cs.toronto.edu>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

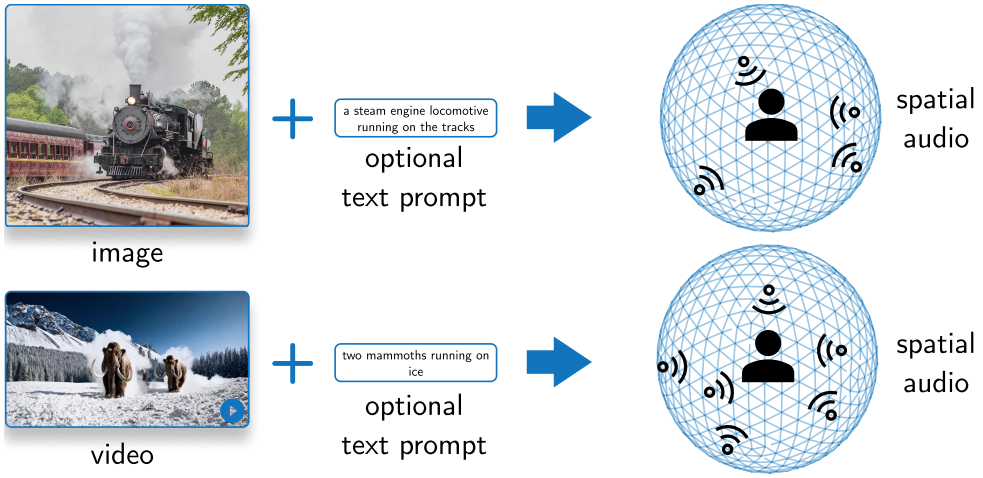

Figure 1: SEE-2-SOUND enables spatial audio creation from real or generated visual content. Our method takes in an image or a sequence of frames (video) and generates spatial audio with a variable number of different sources. The generated audios could also be conditioned on a text prompt.

the 3D location on the viewing sphere for a visual element; (3) generating audio from the regions of interest, and (4) simulating these multiple mono-audio outputs to generate spatial audio. We then explicitly compose these component models to generate spatial audio conditioned on the visual input (e.g. images, animated images (GIFs), or videos) as shown in Figure 1. As our method is compositional and zero-shot, we hope to see SEE-2-SOUND widely used or adapted for content creation and enhancement as well as accessibility.

By generating spatial audio that accurately corresponds to visual elements, we can significantly enhance the immersion of images and videos, creating more realistic and engaging user experiences with digital content. SEE-2-SOUND presents a step towards complete video and image generation, by also incorporating non-verbal audio that accompanies visual generation models (Singer et al., 2023a; Brooks et al., 2024; Blattmann et al., 2023a;b; Luo et al., 2023; Wang et al., 2024). Additionally, our pipeline can also be employed on real visual content.

Our contributions can be summarized as follows.

- We present the challenging new task of generating spatial audio from images or videos. To the best of our knowledge, our work is the first of its kind that tackles this challenging problem.
- We propose SEE-2-SOUND, a novel compositional pipeline that builds on top of and uses off-the-shelf models for segmentation, depth estimation, and audio generation. Due to the nature of our approach, our method is zero-shot in nature and thus is very feasible to use.
- We evaluate SEE-2-SOUND both quantitatively and qualitatively through similarity comparisons and a user study, respectively. To evaluate this method quantitatively we also propose a novel evaluation method for this task.

We do not provide our code since the workshop does not allow submitting supplementary code, however, we intend to to open-source our code and models to foster the reproducibility and extensibility of our work.

2. Methodology

Given a casually captured or generated image, animated image, or video, optionally with a text prompt that describes the content of the media, our objective is to generate spatial audio to accompany the media. We do not impose any qualitative restrictions on the way the media is captured. Our approach, SEE-2-SOUND is a compositional model that has three components (Figure 2): (1) estimating positions for regions of interest (§ 2.1); (2) generating audio for regions of interest (§ 2.2); and (3) generating spatial audio (§ 2.3).

2.1. Source Estimation

We have a visual input, for brevity let $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, represent a (H, W)-resolution and C channel image. Here, we are interested in learning from $\mathbf{I}$, a *set* of 3D coordinates, $\mathbf{x} = (x, y, z)$ representing the positions of regions of interest.

Identifying regions of interest. We first focus on finding out regions of interest from the image, $\mathbf{I}$. We do this by finding a segmentation mask $\mathcal{M} \in \mathbb{R}^{H \times W}$, using a pre-trained

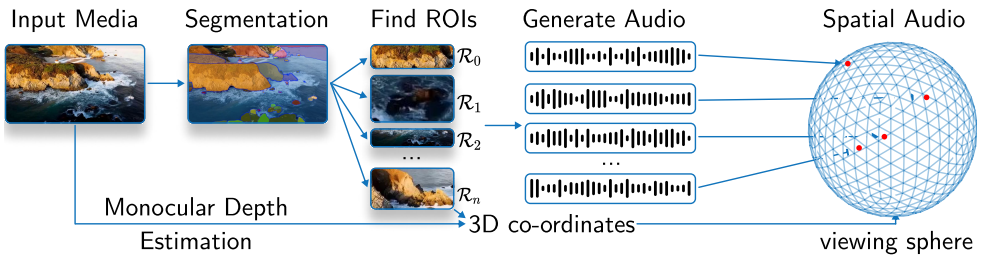

Figure 2: We show an overview of how our approach works. First, we perform image segmentation and find regions of interest for objects in the image. We also perform monocular depth estimation on the image and produce 3D co-ordinate estimating positions for each object in the scene. We then generate mono-audio from the regions of interest. Finally, we place all of these audio according to the 3D co-ordinates on a viewing sphere producing spatial or surround sound.

Segment Anything model (Kirillov et al., 2023). From the segmentation mask $\mathcal{M}$, we extract the contours of each region of interest. Let $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$ represent the set of contours, where each contour $\mathbf{c}_k$ is a sequence of points outlining one region of interest. We then compute the smallest polygon $\mathcal{P}_k$ that bounds each contour $\mathbf{c}_k$. Formally, we solve the following optimization problem for each contour,

$$\min_{\mathcal{P}_k} \text{Area}(\mathcal{P}_k) \text{ , subject to, } \forall \mathbf{c}_k \in \mathcal{P}_k, \tag{1}$$

where $\text{Area}(\mathcal{P}_k)$ denotes the area of the polygon $\mathcal{P}_k$, ensuring that the polygon fully encloses the contour $\mathbf{c}_k$. This allows us to generate a set of minimal bounding polygons, $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k\}$.

We can also modify this to instead set each polygon to an axis-aligned bounding box (AABB) $\mathcal{R}_k$ for each contour $\mathbf{c}_k$. The AABB is defined by the minimum and maximum x and y coordinates of the points in $\mathbf{c}_k$. Formally, for each contour $\mathbf{c}_k$, we define $\mathcal{R}_k$ as,

$$\mathcal{R}_k = \left[\min(\mathbf{c}_k^{(x)}), \max(\mathbf{c}_k^{(x)}), \min(\mathbf{c}_k^{(y)}), \max(\mathbf{c}_k^{(y)})\right]. \tag{2}$$

This allows us to generate a set of AABBs, $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_k\}$.

Estimate monocular depth. For the image $\mathbf{I}$, we also estimate the monocular depth map $\mathcal{D} \in \mathbb{R}^{H \times W}$ using a pre-trained Depth Anything model (Yang et al., 2024). Now we can construct a set of image patches $\mathbf{I_p}$, for the regions of interest from either $\mathcal{P}$ or $\mathcal{R}$ to create a set of variably sized patches, and the 3D coordinates (x, y, z), from the image coordinates and the monocular depth coordinates illustrated in Figure 2. We find that using AABBs usually works better for generating spatial audio.

2.2. Audio Generation

We are now interested in learning from the set of image patches $\mathbf{I_p}$ and an optional text prompt $\mathbf{c}$, a *set* of 4D representations, $(x, y, z, \mathbf{A}^{T \times F})$ representing the positions and audio. We generate mono audio clips conditioned on each of the image patches in $\mathcal{R}$ and the text conditioning $\mathbf{c}$ using a pre-trained CoDi model (Tang et al., 2024). We thus have the 4D representation $(x, y, z, \mathbf{A}^{T \times F})$ for each of the individual image patches in $\{\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \ldots\}$.

2.3. Generating 5.1 Surround Sound Spatial Audio

The goal of surround sound spatial audio generation is to generate ambisonic channels $\phi_N(t)$ from different non-spatial audios and positions $(x, y, z, \mathbf{A}^{T \times F})$. We first construct a virtual acoustic environment, specifically a "Shoe-Box" room (Scheibler et al., 2018). We place the listener centrally at the ground level of this room, serving as a reference point for audio projection and perception. For the source and microphone placements, we adhere to the ITU-R BS 775 configuration (Allen & Berkley, 1979), which is essential for aligning our simulations with the standard 5.1 Surround Sound setups. This configuration includes designated positions for the Front-Left (FL), Front-Right (FR), Center (C), Surround-Left (SL), Surround-Right (SR) and Low Frequency Effects (LFE) channels (Series, 2010).

A critical component of our simulation involves computing the Room Impulse Responses (RIRs) using the Image Source Method (ISM) (Allen & Berkley, 1979). We now calculate the impulse response $h(t)$ by accounting for contributions from multiple virtual sources that arise from reflections off the room's boundaries. For a source at position $\mathbf{s}$ and a receiver at position $\mathbf{r}$, the impulse response is expressed as

$$h(t) = \sum_m \alpha_m \delta(t - \tau_m), \tag{3}$$

where $\tau_m = \frac{\|\mathbf{r} - \mathbf{s}_m\|}{c}$ denotes the time delay; c represents the speed of sound, and α_m are the attenuation factors due to reflections. This model effectively captures both the early reflections and the reverberations, which are crucial for rendering a realistic acoustic environment.

After computing the RIRs, we convolve these responses with the original mono-audio tracks to synthesize the spatial audio for each of the 5.1 channels. The convolution integrates spatial effects into the audio, effectively distributing the sound across the designated channels. The synthesized signal for each channel ch is expressed as

$$\phi_{ch}(t) = \mathbf{A}^{T\times F} * h_{ch}(t), \tag{4}$$

where $\mathbf{A}^{T\times F}$ denotes the original audio signal in the time-frequency domain. Our simulation also incorporates the interference among audio sources which is a significant factor that differentiates the outcome from scenarios where fewer sources are used and then combined. Thus, the resulting sound field in a full simulation exhibits unique characteristics that cannot be replicated by simply summing the outputs of simulations with fewer sources. This interference results in spatial audio properties that are intrinsically linked to the arrangement (x, y, z) and the number of sources in the complete setup.

3. Experiments

To the best of our knowledge, no previous techniques address the problem of spatial sound generation, thus we propose a novel quantitative evaluation technique based on marginal scene guidance (Section 3.1) alongside human evaluation (Section 3.2).

3.1. Quantitative Evaluation

We conduct a quantitative evaluation of the scene guidance of generations produced by our model using various audio similarity metrics. To construct this setup, we first sample images from Laion400M (Schuhmann et al., 2022) and the web. We then generate audio from these images using our method as well as CoDi (Tang et al., 2024) which serves as a baseline. We then use the audio outputs and images to generate modified scene-guided audio generated using AViTAR (Chen et al., 2022a). The use of an unconventional model-based evaluation stems from the lack of existing evaluations or any strong baselines. We report the average audio similarity scores between multiple such generated audio in Table 1. In particular, SEE-2-SOUND consistently generates spatial audio that matches the scene's content as shown by AViTAR (Chen et al., 2022a) only slightly altering the generated audios.

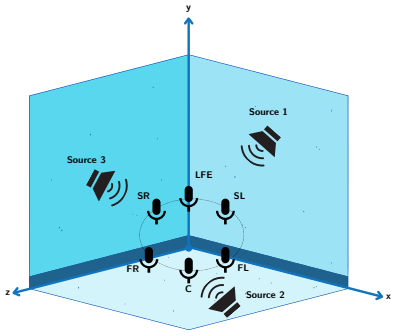

Figure 3: **Overview of the simulation.** We show a schematic of the virtual "ShoeBox" room utilized in our 5.1 surround sound simulation.

3.2. Human Evaluation

Due to the absence of traditional metrics, we also use human evaluation to evaluate our approach. To promote future research in this field, we will open source the evaluation protocol for human evolution. We conduct human evaluations by surveying 43 human evaluators, for which we use randomly selected images and videos from the internet that are non-iconic in nature. We use this qualitative evaluation technique specifically due to the absence of any baseline methods for spatial generation. Additionally, the task of spatial audio generation is challenging since there are multiple possible ground truth spatial audios for a given image or video. This makes it difficult for us to use any kind of similarity metric. Existing similarity metrics such as the LPIPS (Zhang et al., 2018) or CLIP score (Park et al., 2021) and their audio equivalents that make use of embeddings are also not feasible since we do not train embeddings directly for the spatial generation problem.

More details on the tasks in our user study can be found in Appendix F. We evaluate the results for each of these individual subtasks and questions and conducted χ^2 tests to evaluate significance at the $p < 0.05$ level.

4. Conclusion

Our work builds an audio generation pipeline, SEE-2-SOUND, that can generate spatial audio from images, dynamic images, or videos. We achieve this by framing it as a compositional problem and propose a novel framework to use off-the-shelf methods for this challenging new task. To the best of our knowledge, our work is also the first to tackle the problem of spatial audio generation. Evaluating approaches for this novel task is very challenging thus apart from doing a user study we also propose a novel quantitative evaluation method based on measuring scene guidance from audios. We and our human evaluators observe compelling outputs for images and videos both from the internet and produced by generative models. Our method is also zero-shot in nature which makes our method significantly easy to use and extensible. To this extent, we also envision users being able to swap out the models we use for advances and other models in each of the individual sub-problems in our pipeline.

Our work could potentially be used for accessibility applications. The ability to generate spatial audio from visual content could potentially enhance the experience of visual media for individuals with visual impairments. Furthermore, this technology may have applications in medical settings, such as in neurology for stroke patients, and in broader accessibility and rehabilitation contexts.

We hope SEE-2-SOUND and the future work we inspire leads to the generation of truly immersive digital content. We also look forward to seeing other research on the problem we describe in this work. Our codes, models, and results are open-sourced to facilitate future research in this area.

Acknowledgements

The authors thank Bhavya Bhatt of the University of Toronto for his help on surveying the existing literature. This research was enabled in part by the support provided by the Digital Research Alliance of Canada [1]. This research was supported in part with Cloud TPUs from Google's TPU Research Cloud (TRC)[2]. The resources used to prepare this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute [3].

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Allen, J. B. and Berkley, D. A. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Arandjelovic, R. and Zisserman, A. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 435–451, 2018.

Aytar, Y., Vondrick, C., and Torralba, A. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

Aytar, Y., Vondrick, C., and Torralba, A. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., and Rombach, R. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, June 2023b.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409.

Bresin, R., de Witt, A., Papetti, S., Civolani, M., and Fontana, F. Expressive sonification of footstep sounds. *Proceedings of ISon*, 2010:3rd, 2010.

Broderick, J., Duggan, J., and Redfern, S. The importance of spatial audio in modern games and virtual environments. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*, pp. 1–9. IEEE, 2018.

Brooks, T., Peebles, B., Homes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Chen, C., Gao, R., Calamia, P., and Grauman, K. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18858–18868, 2022a.

Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., Calamia, P., Batra, D., Robinson, P., and Grauman, K. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022b.

Chowdhury, S., Ghosh, S., Dasgupta, S., Ratnarajah, A., Tyagi, U., and Manocha, D. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF*

[1] https://alliancecan.ca/

[2] https://sites.research.google/trc/about/

[3] https://vectorinstitute.ai/partnerships/current-partners/

International Conference on Computer Vision, pp. 7884–7896, 2023.

Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., and Wu, Y. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250. IEEE, 2021.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Dickins, G. and Kennedy, R. Towards optimal soundfield representation. In *Audio Engineering Society Convention 106*. Audio Engineering Society, 1999.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Gan, C., Zhao, H., Chen, P., Cox, D., and Torralba, A. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7053–7062, 2019.

Gan, C., Huang, D., Chen, P., Tenenbaum, J. B., and Torralba, A. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 758–775. Springer, 2020.

Gao, R. and Grauman, K. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 324–333, 2019.

Garg, R., Gao, R., and Grauman, K. Geometry-aware multi-task learning for binaural audio generation from video. In *British Machine Vision Conference*, 2021.

Gerzon, M. A. Periphony: With-height sound reproduction. *Journal of the audio engineering society*, 21(1):2–10, 1973.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, oct 2020. ISSN 0001-0782. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

Hu, D., Li, X., et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3574–3582, 2016.

Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., and Dou, D. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33: 10077–10087, 2020.

Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023a.

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., and Zhao, Z. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/huang23i.html.

Iashin, V. and Rahtu, E. Taming visually guided sound generation. In *The 32st British Machine Vision Virtual Conference*. BMVA Press, 2021.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., and Girshick, R. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Kirkeby, O., Nelson, P. A., Orduna-Bustamante, F., and Hamada, H. Local sound field reproduction using digital signal processing. *The Journal of the Acoustical Society of America*, 100(3):1584–1593, 1996.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CYK7RfcOzQ4.

Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., and Kumar, K. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36, 2024.

Lam, M. W., Tian, Q., Li, T., Yin, Z., Feng, S., Tu, M., Ji, Y., Xia, R., Ma, M., Song, X., et al. Efficient neural music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Li, D., Langlois, T. R., and Zheng, C. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.

Liang, S., Huang, C., Tian, Y., Kumar, A., and Xu, C. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36, 2024.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, June 2023.

Lin, Y.-B. and Wang, Y.-C. F. Exploiting audio-visual consistency with partial supervision for spatial audio generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2056–2063, 2021.

Ling, H., Kim, S. W., Torralba, A., Fidler, S., and Kreis, K. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. *arXiv preprint arXiv:2312.13763*, 2023.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. Audioldm: text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 21450–21474, 2023.

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., and Plumbley, M. D. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–13, 2024. doi: 10.1109/TASLP.2024.3399607.

Luo, S., Yan, C., Hu, C., and Zhao, H. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., and Tan, T. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10209–10218, 2023.

Morgado, P., Nvasconcelos, N., Langlois, T., and Wang, O. Self-supervised generation of spatial audio for 360 deg video. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Murphy, D. and Neff, F. Spatial sound for computer games and virtual reality. In *Game sound technology and player interaction: Concepts and developments*, pp. 287–312. IGI Global, 2011.

Owens, A. and Efros, A. A. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 631–648, 2018.

Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., and Torralba, A. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 801–816. Springer, 2016.

Parida, K. K., Srivastava, S., and Sharma, G. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3347–3356, 2022.

Park, D. H., Azadi, S., Liu, X., Darrell, T., and Rohrbach, A. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Ratnarajah, A., Ghosh, S., Kumar, S., Chiniya, P., and Manocha, D. Av-rir: Audio-visual room impulse response estimation. *arXiv preprint arXiv:2312.00834*, 2023.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=piLPYqxtWuA.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., and Torralba, A. Self-supervised audio-visual co-segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2357–2361. IEEE, 2019.

Rumsey, F. *Spatial audio*. Taylor & Francis, 2012.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Scheibler, R., Bezzam, E., and Dokmanić, I. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355. IEEE, 2018.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Senocak, A., Oh, T.-H., Kim, J., Yang, M.-H., and Kweon, I. S. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4358–4366, 2018.

Series, B. Multichannel stereophonic sound system with and without accompanying picture. *International Telecommunication Union Radio-communication Assembly*, 2010.

Sheffer, R. and Adi, Y. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., and Taigman, Y. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=nJfylDvgzlq.

Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., et al. Text-to-4d dynamic scene generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 31915–31929, 2023b.

Son Chung, J., Senior, A., Vinyals, O., and Zisserman, A. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6447–6456, 2017.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Tang, Z., Yang, Z., Khademi, M., Liu, Y., Zhu, C., and Bansal, M. Codi-2: In-context, interleaved, and interactive any-to-any generation, 2023.

Tang, Z., Yang, Z., Zhu, C., Zeng, M., and Bansal, M. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

Tian, Y., Shi, J., Li, B., Duan, Z., and Xu, C. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 247–263, 2018.

Tian, Y., Li, D., and Xu, C. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 436–454. Springer, 2020.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. In *Arxiv*, 2016. URL https://arxiv.org/abs/1609.03499.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/

3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, C., Serrano, A., Pan, X., Wolski, K., Chen, B., Seidel, H.-P., Theobalt, C., Myszkowski, K., and Leimkühler, T. A neural implicit representation for the image stack: Depth, all in focus, and high dynamic range. *ACM Transactions on Graphics*, 42(6), 2023. doi: 10.1145/3618367.

Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., and Zhou, J. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024.

Wu, Y., Zhu, L., Yan, Y., and Yang, Y. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6292–6300, 2019.

Xie, B. Spatial sound-history, principle, progress and challenge. *Chinese Journal of Electronics*, 29(3):397–416, 2020.

Xue, J., Deng, Y., Gao, Y., and Li, Y. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *arXiv preprint arXiv:2401.01044*, 2024.

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., and Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.

Ye, Y., Yang, W., and Tian, Y. Lavss: Location-guided audio-visual spatial audio separation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5508–5519, 2024.

Yu, J., Zhang, S.-X., Wu, J., Ghorbani, S., Wu, B., Kang, S., Liu, S., Liu, X., Meng, H., and Yu, D. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6984–6988. IEEE, 2020.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhang, W., Samarasinghe, P. N., Chen, H., and Abhayapala, T. D. Surround by sound: A review of spatial audio recording and reproduction. *Applied Sciences*, 7(5):532, 2017.

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 570–586, 2018.

Zhou, H., Liu, Y., Liu, Z., Luo, P., and Wang, X. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9299–9306, 2019.

Zhou, H., Xu, X., Lin, D., Wang, X., and Liu, Z. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 52–69. Springer, 2020.

Zhou, Y., Wang, Z., Fang, C., Bui, T., and Berg, T. L. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3550–3558, 2018.

Ziv, A., Gat, I., Lan, G. L., Remez, T., Kreuk, F., Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*, 2024.

A. Related works

The audio generation field has been gaining a lot ofmuch attention lately, primarily in models that create a diverse range of of audio content types. From speech synthesis to music generation and sound effects creation, past efforts have explored numerous aspects of auditory information synthesis. Learning algorithms have led to ever-improving techniques in the models for generating high-quality, diverse audio. This surge in interest has fueled research efforts pushing the boundaries of audio generation.

A.1. Conditional Audio Generation

Recent work in conditional audio generation has made significant progress using latent diffusion models (Liu et al., 2023) such as AudioLDM (2) (Liu et al., 2023; 2024) and Make-An-Audio (Huang et al., 2023b). Audiogen (Kreuk et al., 2023) was one of the first to create a text-to-audio environment, treating audio generation as a conditional language modeling task. Building on this, AudioLDM-2 (Liu et al., 2024) introduces a unified audio generation framework that addresses different objectives for speech, music, and sound effects, using the "language of audio" (LoA) and a GPT-2 (Radford et al., 2019) model for translation and latent diffusion (Rombach et al., 2022) for generation.

AudioLM (Borsos et al., 2023) employs neural codecs (Wang et al., 2023) for long-term consistency in audio generation, mapping input audio to discrete tokens, and using a language modeling task. Meanwhile, DiffSound (Yang et al., 2023), using a text encoder, decoder, vector-quantized variational autoencoder (VQVAE) (Kingma & Welling, 2013), and "vocoder", introduces a mask-based text generation strategy (MBTG) to generate text descriptions from audio labels to address the scarcity of audio-text paired data. NaturalSpeech (Tan et al., 2024) integrates a memory component in variational autoencoders (Kingma & Welling, 2013) to achieve human-level quality in text-to-speech systems.

Significant progress in text-to-music generation has been made by notable models such as MusicLM (Agostinelli et al., 2023), Noise2Music (Huang et al., 2023a), MusicGen (Copet et al., 2024) and MeLoDy (Lam et al., 2024). MusicLM (Agostinelli et al., 2023) uses contrastive pretraining to align music and language embeddings, improving performance without explicit text using a semantic modeling stage based on w2v-BERT (Chung et al., 2021). MusicGen (Copet et al., 2024) adopts a language modeling approach with melodic conditioning, enhancing control over output. MeLoDy (Lam et al., 2024) employs dual-path diffusion-based modeling for computational efficiency in music generation.

The image-to-audio and video-to-audio domains have also seen significant advancements. For instance, Im2Wav (Sheffer & Adi, 2023) generates audio from images or image sequences, using Transformers (Vaswani et al., 2017) to manipulate hierarchical discrete audio representation derived from VQ-VAE (Van Den Oord et al., 2017). Similarly, based on GANs (Goodfellow et al., 2020): SpecVQGAN (Iashin & Rahtu, 2021) enhances visually driven audio synthesis from videos, while RVQGAN (Kumar et al., 2024) achieves high-fidelity audio compression. (Aytar et al., 2017) introduces aligned representations that help in cross-modal retrieval and transferring between audio, visual, and text modalities. Furthermore, AV-NeRF (Liang et al., 2024) synthesizes new videos with spatial audio along different camera trajectories, and AV-RIR (Ratnarajah et al., 2023) estimates audio-visual room impulse responses from images and audio.

Multi-modal generative models of note include CoDi(-2) (Blattmann et al., 2023b; Tang et al., 2023). CoDi (Blattmann et al., 2023b) generates multiple output modalities from various input combinations using a composable generation strategy. CoDi-2 (Tang et al., 2023) extends this with enhanced modality-language alignment for encoding and generative tasks, improving zero-shot capabilities in multi-modal generation, reasoning, and interactive conversations.

Recent works like MAGNeT (Ziv et al., 2024) introduce a masked generative sequence modeling method for direct audio token generation, achieving higher efficiency in text-to-music and text-to-audio tasks. DIFF-FOLEY (Luo et al., 2024) improves the quality, relevance, and synchronization of video-to-audio synthesis using latent diffusion and contrastive audiovisual pre-training. Affusion (Xue et al., 2024) enhances text-to-audio generation by leveraging text-to-image diffusion models, producing high-quality audio closely aligned with text prompts. These advancements significantly expand the possibilities of audio generation.

A.2. Acoustic Matching

Recent advancements in acoustic matching integrate visual and audio cues to enhance realism and immersion in generated content. AViTAR (Chen et al., 2022a) performs visual-acoustic matching using a cross-modal transformer to integrate visual properties into audio and learn from in-the-wild videos. AdVerb (Chowdhury et al., 2023) uses a geometry-aware

cross-modal transformer to enhance reverberation by leveraging visual cues, significantly outperforming traditional methods. SoundSpaces 2.0 (Chen et al., 2022b) provides geometry-based audio rendering for 3D environments, supporting tasks like audio-visual navigation and source localization with high fidelity and realism.

A.3. Construction of Surround Sound

In 1881, Ader used a pair of microphones and telephone receivers to relay sound signals at the Paris Opera (Rumsey, 2012), a first-of-its-kind spatial (stereophonic) sound demonstration. In the past three decades, the development of digital signal processing (Kirkeby et al., 1996), computer and internet technologies has rapidly evolved spatial sound (Xie, 2020). Together, they have made spatial audio commonplace now (Zhang et al., 2017).

In the context of virtual environments generated by game engines, spatial audio is crucial to provide an immersive experience (Broderick et al., 2018). Traditionally, sounds in virtual worlds are created by recording individual sound sources separately and then blending them according to the intended scene setup (Murphy & Neff, 2011) using various audio processing techniques (Gerzon, 1973; Dickins & Kennedy, 1999). But this recording process is costly (Zhang et al., 2017) and optimizing it involves generating spatial audio. While spatializing mono-audio is a popular area of research (discussed below), to the best of our knowledge, there have not been any approaches to generate spatial audio directly.

A.4. Audio Spatialization

Recent advances in audio spatialization have focused on the integration of audio and visual modalities to enhance realism. Sep-Stereo (Zhou et al., 2020) uses an associative pyramid structure to generate stereo output by combining audio and visual information. Similarly, LAVSS (Ye et al., 2024) introduces a location-guided audio-visual spatial audio separator, highlighting the importance of visual cues in audio generation.

Several studies have synchronized visual cues and encoders to produce natural sound for dynamic video scenarios (Zhou et al., 2018). Methods like those in (Gan et al., 2019; Gao & Grauman, 2019) use a U-Net to encode monaural input and decode binaural output with visual guidance.

Other approaches, such as (Gan et al., 2020; Kreuk et al., 2023) generate audio samples conditioned on various inputs such as text, motion key points, and position information (Garg et al., 2021; Parida et al., 2022), respectively. Techniques like (Morgado et al., 2018) utilize end-to-end trainable neural networks for sound source separation and localization in 360-degree videos. Furthermore, works like (Li et al., 2018) synthesize directional impulse responses by exploiting the spatial and directional characteristics of the reverberation of the room.

A.5. Audio Visual Learning

In recent years, there has been a growing interest in audio-visual learning, focusing on understanding the inherent connection between visual and auditory stimuli within vision research. Noteworthy works have delved into explaining the natural synchronization between visual and audio elements (Arandjelovic & Zisserman, 2017; Aytar et al., 2016; Owens et al., 2016). This approach finds application in diverse domains, including speech recognition utilizing both audio and visual cues (Son Chung et al., 2017; Hu et al., 2016; Yu et al., 2020; Zhou et al., 2019), event localization through joint analysis of audio and visual information (Tian et al., 2018; 2020; Wu et al., 2019), and sound source localization (Arandjelovic & Zisserman, 2018; Hu et al., 2020; Rouditchenko et al., 2019; Senocak et al., 2018; Tian et al., 2018; Zhao et al., 2018). Additionally, there have been explorations in the use of audio-visual data for unsupervised representation learning, as demonstrated in (Owens & Efros, 2018; Owens et al., 2016).

In contrast to these modalities, our focus diverges as we undertake a distinct task: the direct generation of surround sound spatial audio from images.

B. Preliminary: Conditional LDM

Latent Diffusion models (LDMs) (Rombach et al., 2022) project input images into a lower-dimensional latent space. Unlike traditional diffusion models that operate in a high-dimensional pixel space, LDMs perform forward and backward diffusion processes in a lower-dimensional learned latent space. For generating audio, we have a conditional LDM $\boldsymbol{\epsilon}_\theta$, which can be interpreted as a sequence of denoising autoencoders $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \ldots); t \sim \mathcal{U}(1, \ldots, T)$, trained using a squared error loss

Table 1: **Quantitative Evaluation Results.** We report average similarity scores between different audios. These scores for a method (row) are computed between audio produced by said method and that when enhanced by AViTAR (Chen et al., 2022a). Lower MFCC-DTW scores and higher ZCR, Chroma, and Spectral scores indicate higher similarity.

(+ AViTAR (Chen et al., 2022a))	**MFCC-DTW** $\downarrow$	**ZCR** $\uparrow$	**Chroma** $\uparrow$	**Spect** $\uparrow$
CoDi (Tang et al., 2024) (baseline)	0.0008	0.80	0.70	0.85
SEE-2-SOUND (ours)	0.00003	0.95	0.77	0.95

Table 2: **Human Evaluation Results.** We evaluate our model performance across metrics identified through a user study. Our hypothesis was that the responses follow a uniform distribution. For the quality questions, we considered scores on a scale of 1-5 and scores above 3 to be correct. For the distance and localization problems, we used the patches and coordinates generated by our approach as correct answers. For adversarial identification and matching problems, their nature provided ground-truth correct answers. The significance of human preference is evaluated with χ^2-tests at $p < 0.05$.

Metric	*p-value* ($\downarrow$)
Spatial Audio Quality	
Realism	4.174×10^{-5}
Immersion	1.036×10^{-4}
Accuracy	1.036×10^{-4}
Clarity	5.927×10^{-6}
Consistency	2.089×10^{-6}
Audio Identification	
Overall Localization	4.523×10^{-4}
Audio Direction Identification	2.880×10^{-7}
Distance Identification	3.103×10^{-2}
Adversarial Identification	9.345×10^{-2}
Audio Matching	7.431×10^{-5}

(representing the denoising term in the ELBO) to denoise some variably noisy audio signals using the following objective,

$$\mathbb{E}_{\mathcal{E}(\mathbf{x}),\mathbf{c},\boldsymbol{\epsilon}\sim\mathcal{N}(0,I),t}\left[|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}))|_2^2\right], \tag{5}$$

where $\mathbf{x}_t$ is some noisy audio signal; $\mathbf{c}$ is a conditioning vector; τ_θ is an encoder that projects $\mathbf{c}$ to an intermediate representation; $\mathcal{E}$ encodes $\mathbf{x}_t$ into a latent representation $\mathbf{z}_t$; and we jointly optimize $\boldsymbol{\epsilon}_\theta$ and τ_θ.

C. Quantitative and User Study Results

We demonstrate our results in Tables 1 and 2. We particularly show in Table 1 that our approach consistently generates spatial audio that matches the scene's content as shown by AViTAR (Chen et al., 2022a) only slightly altering the audio generated by our approach.

we demonstrate our output in the supplementary, from which one can observe that SEE-2-SOUND produces compelling spatial audio that aligns well with videos and images which are either generated from video generation models or taken from the Internet. We perform user testing on 43 human evaluators and report the metrics for our human evaluation in Table 2. The users indicated a statistically significant preference for SEE-2-SOUND in terms of spatial audio quality, direction of sound, and audio quality.

D. Discussion

Our work introduces a method for generating spatial sound from images and videos. We envision many avenues for potential impact based on our technique, including applications related to enhancing images and videos generated by learned methods, making real images interactive, human-computer interaction problems, or accessibility.

Limitations. Although SEE-2-SOUND produces compelling spatial audio clips, there are several limitations and avenues for future work. First, our framework may not detect all of the fine details in images and videos, nor does it produce audio for every fine detail. If we try to produce audio for many fine structures with this framework using upscaling or other zooming-in methods, we observe a lot of interference within the audios.

Second, SEE-2-SOUND does not generate audios based on any motion cues but solely based on the image priors. We could benefit from having separate appearance and motion backbones, which may potentially allow us to use any motion information to refine the generated audio.

Third, currently, our method is not able to work in real time on a Nvidia A100-80 GPU on which the experiments were conducted. However, since SEE-2-SOUND relies on composing off-the-shelf models, we could easily swap out the models we use for other models that solve the individual subproblems. We can easily incorporate advances in any of the subproblem areas within our framework to potentially get to real-time capabilities.

Lastly, inherent in the capabilities of generative models such as SEE-2-SOUND, which excel in producing synchronized and realistic outputs, is a significant concern aboutregarding the potential generation and dissemination of deepfakes. The exploitation of this technology by malicious entities poses a notable risk even though we do not generate speech, as they could fabricate highly convincing counterfeit audio clips. We condemn any such application and hope to be able to work on safeguards for such approaches in the future.

E. Details of our Model

We provide details of our approach with which we demonstrate all of our experiments and results. Our experiments ran on a single A100-80 GPU. Running our approach on an image takes $\sim 12 - 15$ seconds (non-initial runs; initial runs can take up to ~ 3 minutes) for images and ~ 5 minutes on videos, which are less than 20 seconds like the ones we experiment with. Our method can be easily swapped for other off-the-shelf models, however we limit our experiments to these settings. Our human evaluation experiments were also performed without any text prompts to solely determine our framework's performance.

Source Estimation. For source estimation we use a Segment Anything (Kirillov et al., 2023) model. We use the ViT-H/16 (Dosovitskiy et al., 2021) with 14×14 windowed attention as our image encoder. Following standard approaches we rescale any inputs to 1024×1024 by rescaling and padding. For the decoder we use the same design and weights as used by Segment Anything (Kirillov et al., 2023). For monocular depth estimation we use a Depth Anything (Yang et al., 2024) model. We use ViT-L/14 (Dosovitskiy et al., 2021) as our image encoder.

Audio generation. For audio generation we use a CoDi (Tang et al., 2024) model and use the Video LDM, Audio LDM, and Text LDM components. For inference, we sample the model through 500 diffusion steps.

Generating 5.1 Surround Sound. We use a sampling rate of 16000 Hz for audio processing. For simulating the acoustic environment, we utilize a "ShoeBox" room model from Pyroomacoustics (Scheibler et al., 2018). The dimensions of the room are dynamically calculated based on the size of the image. Specifically, the dimensions width $w =$ image_width, height $h =$ image_height, and length $l = 0.5\cdot$ image_width. In our virtual room, the microphones are placed at the following coordinates for each channel,

- Front-Left (FL): [0, 0, 100]
- Front-Right (FR): [0, 0, -100]
- Center (C): [100, 0, 0]
- Low-Frequency Effects (LFE): [0, -100, 0]
- Surround-Left (SL): [-100, 0, 0]
- Surround-Right (SR): [0, 100, 0]

F. User Study Protocol

We now describe our human evaluation technique. Our human evaluation task has the following key components.

Spatial Audio Quality Human evaluators qualitatively rate the realism, immersion, and accuracy of two images or videos with spatial audio generated by our method. These aspects are measured using a semantic differential scale.

Direction Identification Evaluators estimate the sensation of direction and distance of an object given the generated spatial audio. This assesses the effectiveness of the spatial sound in conveying the sensation of sound originating from multiple directions.

Adversarial Identification Evaluators are randomly given a mix of generated spatial audio and mono audio. They are tasked with identifying whether the given audio is indeed spatial audio, and their accuracy is measured.

Audio Quality Evaluators are provided with three clips of spatial audio and three images or videos. They perform a matching task to pair the audio clips with the corresponding visual content, and their matching accuracy is measured.

We demonstrate the user study that the participants were given. We however removed any YouTube video links we had as well as a consent form and institutional details here.

F.1. Consent Form

Your decision to participate is voluntary, and you are free to withdraw at any time without any consequences. You are not required to answer any questions that you do not feel comfortable answering.

Please ensure that you understand the following:

1. The purpose of the survey is to evaluate AI for spatial audio generation from image inputs and optional text prompts for the purpose of improving the quality of the outputs from these systems.
2. I will be asked a questionnaire consisting of comparisons between pairs of images, audios, and videos.
3. The questionnaire takes approximately 15 minutes to complete, and will be conducted online.
4. I may withdraw from the questionnaire at any time by exiting the questionnaire screen. If I choose to withdraw, all partial data collected will be discarded.
5. The researchers do not foresee any risks or stresses on the participant beyond what one might experience in a typical day of work.
6. The researchers also do not foresee any immediate benefits for the participant beyond what one might experience in a typical day of work.
7. The results collected will help researchers develop systems for using AI models to generate spatial audio from image inputs and optional text prompts.
8. I agree to have my responses documented and recorded digitally.
9. All data collected from the questionnaire will be kept secure. Only the researchers involved in this study will have access to the information I provide. All data collected will be destroyed 5 years after publication of results.
10. Data collected may be used in research journals, conferences, or other scholarly activities. In these cases, an anonymous subject identifier will be used and no identifying information will be provided to the audience.
11. I am at least 18 years of age.
12. I am free to ask questions about the process at any time. I can ask questions in person, or by contacting through email.
13. If requested, I will receive a copy of this form for my records.

The research study you are participating in may be reviewed for quality assurance to make sure that the required laws and guidelines are followed. If chosen, (a) representative(s) of the Human Research Ethics Program (HREP) may access study-related data and/or consent materials as part of the review. All information accessed by the HREP will be upheld to the same level of confidentiality that has been stated by the research team. If you have any questions about your rights as a participant, please contact the Ethics Review Office at (this email) or (this phone number).

F.2. Part 1: Familiarization

Before we begin, please note that the optimal testing condition involves 5.1 surround speakers. However, for this test, we ask participants to use headphones.

Watch this preselected video with correct spatial audio to familiarize yourself with the spatial audio experience.

F.3. Part 2: Rating Spatial Audio with Images

You will be exposed to images paired with corresponding spatial audio. Some of these may be generated by our system.

Rate the realism, immersion, accuracy, clarity, and consistency of the audio representation in relation to the visual content on a scale of 1 to 5, where 1 is not realistic/immersive/accurate/clear/consistent, and 5 is very realistic/immersive /accurate/clear/consistent.

Here is how we have defined these terms:

- **Realism:** How authentically the spatial audio mirrors real-world sound distribution.
- **Immersion:** The extent to which spatial audio engages and envelops the listener.
- **Accuracy:** The precision of alignment between spatial audio and visual content.
- **Clarity:** The degree to which individual sound elements are distinct and perceptible.
- **Consistency:** The uniformity of spatial audio representation across different images.

Video 1
Please rate the spatial audio in Video 1 on the following parameters, where 1 is not realistic/immersive/accurate/clear/consistent, and 5 is very realistic/immersive /accurate/clear/consistent.

Video 2
Please rate the spatial audio in Video 2 on the following parameters, where 1 is not realistic/immersive/accurate/clear/consistent, and 5 is very realistic/immersive /accurate/clear/consistent.

F.4. Part 3: Direction and Distance Identification

You will be exposed to videos paired with corresponding spatial audio. Some of these may be generated by our system.

Identify the direction and distance of sound sources within the audio landscape as depicted in the videos.

Video 3

1. Where did you perceive the sound source for the trumpet to be located in the audio clip? (You can choose multiple options)
    - Left
    - Right
    - Center
    - Behind
    - In front
    - Above
    - Below

2. How would you describe the distance of the sound source for the trumpet in the audio clip?
   - Very Close
   - Moderately Close
   - Moderately Far
   - Very Far
3. Where did you perceive the sound source for the drums to be located in the audio clip? (You can choose multiple options)
   - Left
   - Right
   - Center
   - Behind
   - In front
   - Above
   - Below
4. How would you describe the distance of the sound source for the drums in the audio clip?
   - Very Close
   - Moderately Close
   - Moderately Far
   - Very Far
5. How accurately could you localize different sound sources within the audio landscape?
   - Very Accurate
   - Moderately Accurate
   - Not Accurate

F.5. Part 4: Real vs. Fake Identification

You will be shown 2 clips, one of them has spatial audio generated by and another one has mono audio of the same clips. You will be asked to identify which of them is ”real” spatial audio and ”fake” spatial audio.

Video 4a

Video 4b

Which of the clips is more likely to be real spatial audio?

- Video 4a
- Video 4b

F.6. Part 5: Matching Images to Spatial Audio

You will be shown three images and corresponding spatial audio. Some of these may be generated by our system. You will be asked to match each spatial audio clip to its corresponding image.

Image 1 was shown.

Spatial Audio 1

Image 2 was shown.

Spatial Audio 2

Image 3 was shown.

Spatial Audio 3

Match each spatial audio clip to its corresponding image.

G. Quantitative Evaluation Metrics

The evaluation metrics we utilized are listed below:

1. **MFCC-DTW Similarity**: Measures the similarity between the Mel-Frequency Cepstral Coefficients (MFCCs) using Dynamic Time Warping (DTW) distance.
2. **Zero-Crossing Rate (ZCR) Similarity**: Measures the similarity between the zero-crossing rates of the audio files.
3. **Rhythm Similarity**: Calculates the similarity between the rhythm patterns using Pearson correlation coefficient.
4. **Chroma Similarity**: Measures the similarity between the chroma features.
5. **Spectral Contrast Similarity**: Compares the spectral contrast of the audio files.
6. **Perceptual Similarity**: Quantifies the perceptual similarity using the Short-Time Objective Intelligibility (STOI) metric.

We excluded the perceptual and rhythm similarity metrics for the following reasons:

- **Perceptual Similarity**: Comparing spatial audio to mono audio using perceptual similarity (such as the Short-Time Objective Intelligibility (STOI) metric) is not meaningful, as the perceptual qualities are inherently different.
- **Rhythm Similarity**: Our framework does not generate music; hence, rhythm similarity is irrelevant in our context.

H. Licenses

Here we list the licensing information for code bases, pre-trained models, and data used:

- Segment Anything (Kirillov et al., 2023) (code and weights): Apache License 2.0, https://github.com/facebookresearch/segment-anything/tree/main
- Depth Anything (Yang et al., 2024) (code and weights): Apache License 2.0, https://github.com/LiheYoung/Depth-Anything/tree/main
- CoDi (Tang et al., 2024) (code and weights): MIT License, https://github.com/microsoft/i-Code/tree/main/i-Code-V3
- Visual Acoustic Matching (Chen et al., 2022a) (code and weights): Creative Common CC-BY-NC Licence, https://github.com/facebookresearch/visual-acoustic-matching
- Laion400M (Schuhmann et al., 2022): Creative Common CC-BY 4.0 License, https://laion.ai/blog/laion-5b/
- Images from web: Creative Common CC-BY 4.0 License

Our code and assets are under the Apache 2.0 License.