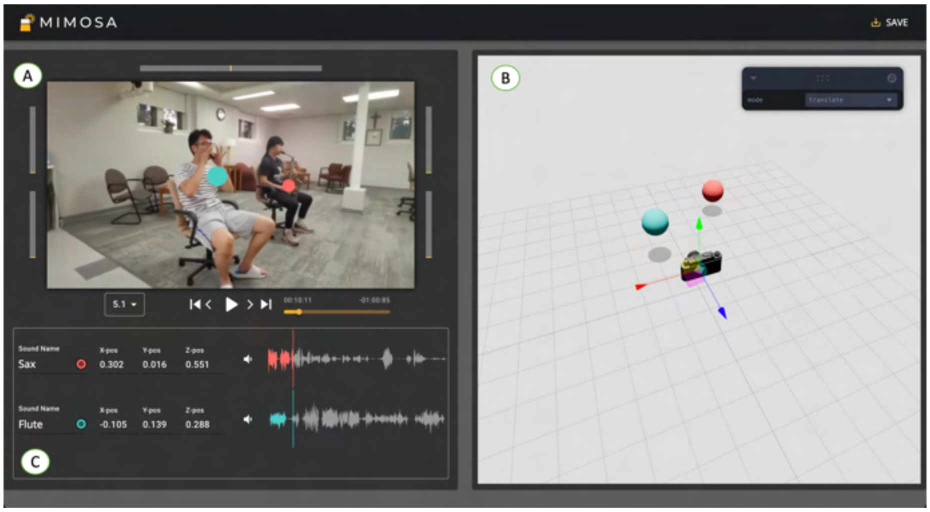
# MIMOSA: Human-in-the-Loop Generation of Spatial Audio from Videos with Monaural Audio

Zheng Ning[1#], Zheng Zhang[1#], Jerrick Ban[1], Kaiwen Jiang[1,2],
Ruohong Gan[1,3], Yapeng Tian[4*], and Toby Jia-Jun Li[1*]

[1]University of Notre Dame [2]Beijing Jiaotong University [3]Sichuan University
[4]University of Texas at Dallas
{zning,zzhang37,jban,kjiang2,rgan,toby.j.li}@nd.edu,
yapeng.tian@utdallas.edu



| Sound Name | | X-pos | Y-pos | Z-pos |
|---|---|---|---|---|
| Sax | ○ | 0.302 | 0.016 | 0.551 |
| Flute | ○ | -0.105 | 0.139 | 0.288 |

**Fig. 1.** The main interface of Mimosa.

## 1 Introduction

The design of audio is half of the story when creators make video content. In particular, spatial audio has been shown to significantly contribute to improving viewers' memory, understanding, and engagement of the video content [1,4,16]. However, the cost and learning barrier of spatial audio recording equipment [14, 15] and the difficulty of post-processing [2,12] limit the wide adoption of spatial audio in the video creation of amateur content creators.

While end-to-end machine learning (ML) models such as [5,13,20] have shown significant progress in generating spatial audio from videos with monaural audio, several limitations prevent them from completely fulfilling the actual needs of content creators: First, despite recent progress, these models only achieve limited accuracy while the end-to-end "black-box" nature of these models prevent users from easily validating their results and fix errors. Second, the model

[#] The first two authors equally contributed to the work.

[*] Corresponding authors.

can only give predictions for the "ground truth", while the ideal content creation process is an artistic expression [3,21] that goes beyond the ground truth. Therefore, we argue that this is not only an ML challenge but more importantly, a human-computer interaction (HCI) challenge where we need to empower content creators to collaborate with computational tools.

To address these issues, we introduce MIMOSA[1](Fig. 1), an interactive human-AI collaborative tool that helps amateur content creators create immersive spatial audios for videos with monaural or stereo audios.

The design of MIMOSA exemplifies a human-in-the-loop human-AI collaboration approach that, instead of using the state-of-art end-to-end "black-box" machine learning model, employs a multi-step pipeline carefully designed to align with the human user's cognitive and perceptual workflow. MIMOSA's multi-step pipeline (Fig. 2) consists of object detection, depth estimation, soundtrack separation, audio tagging, and spatial audio rendering modules. While the prediction accuracy of this pipeline *alone* might compare with the state-of-art end-to-end models, modules in this pipeline produce useful intermediate results that allow user understanding, validation, repairing, manipulation, and recreation at each step. Therefore, its *human-model-collaborative* performance could be better than the state-of-art model. An interactive direct-manipulation [7] interface designed with mixed-initiative interaction principles [6] allows content creators to make sense of model outputs and provide inputs in a manner that is both familiar and natural to users and useful for the models. Furthermore, the enhanced user control in the process would also allow greater *expressiveness* for users to "go above the groundtruth" to achieve their desired effects, which is particularly useful for AI-enabled content creation tools. To promote the real-world deployment, we also implemented an extension of MIMOSA for Adobe Premiere Pro[2].

## 2 A Walkthrough of an Example Use Case

To illustrate the usage of MIMOSA, we demonstrate an example use case where the user already has a video with monaural sound of a duet band playing in a room, and the user wishes to augment the video with spatial audio effects.

The user starts Adobe Premiere Pro, loads a video, and selects the target video clips. From the "Extension" menu, the user launches MIMOSA. The MIMOSA processing pipeline (Section 3.1) will run in the backend to separate the audio into individual soundtracks, predict the tag for each soundtrack, identify and localize the sounding object, and estimate the depth of each sounding object. The results of the pipeline will be visualized in the video previewing panel (Fig.1A) and the 3D direct manipulation interface (Fig.1B). Using these features, the user can view the predicted sounding objects and make any edits if needed.

In the video previewing panel, several dots in different colors will be shown in an overlay on top of the video playback to indicate the inferred positions of sounding objects. If the inferences are inaccurate, the user may directly drag

[1] MIMOSA is an acronym for **M**agnifying **I**mmersiveness by **M**anipulating **O**bjects in **S**patial **A**udio

[2] A demo video is available at `https://www.youtube.com/watch?v=3X3gkDkW9bc`.

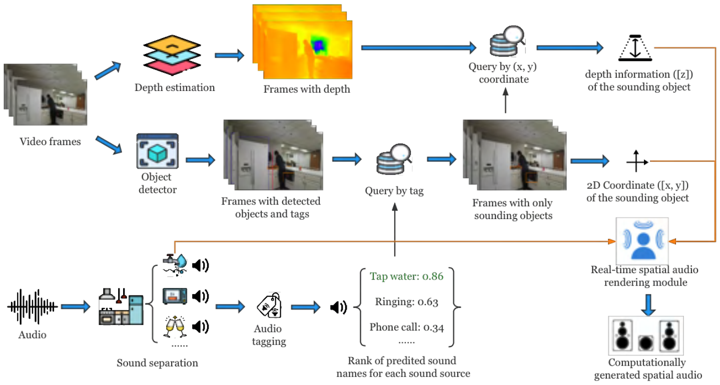

Video frames
Depth estimation
Frames with depth
Query by (x, y) coordinate
depth information ([z]) of the sounding object
Object detector
Frames with detected objects and tags
Query by tag
Frames with only sounding objects
2D Coordinate ([x, y]) of the sounding object
Real-time spatial audio rendering module
Computationally generated spatial audio
Audio
Sound separation
Audio tagging
Tap water: 0.86
Ringing: 0.63
Phone call: 0.34
......
Rank of predited sound names for each sound source

**Fig. 2.** MIMOSA's visual object oriented audio spatialization pipeline

these dots to make them align with the actual sounding objects. At the same time, a 3D direct manipulation interface will visualize their positions in a simulated 3D space, allowing users to adjust the positions at a finer granularity. The user can select an output format (e.g., quadraphonic sound, 5.1 channels) and the surround sound with spatial effect will be rendered in real-time for preview based on the positions of each sounding object. The loudness of each channel will also be visualized to indicate the output of each channel.

Besides predicting the actual positions of sounding objects for generating realistic spatial audio effects that *resemble* the ground truth, content creators may also "go beyond the ground truth" to author customized spatial audio effects with MIMOSA. For example, the content creator can manipulate the location and orientation of the reference point by directly manipulating the camera recorder in the 3D interface. Instead of having the band play in front of the viewer, the content creator may move the viewing point to the back of the band, near the saxophone player, or between the two players. They may also manipulate the positions of sounding objects so they may, for example, move around the viewer or out of sight towards the back of the viewer. When the content creator is done with editing, they may click on the "SAVE" button so MIMOSA will export the generated spatial audio and import it back into Adobe Premier Pro.

## 3 The MIMOSA System

In this section, we illustrate MIMOSA's processing pipleine and its key interaction strategies for facilitating effective human-AI collaboration.

### 3.1 A Visual Object Oriented Audio Spatialization Pipeline

We propose a visual object oriented audio spatialization pipeline (Fig.2) for MIMOSA. In the pipeline, we first discover sounding objects in video frames and separate individual sound sources in audio, then generate spatial audio effects for each separated sound associated with the corresponding sounding object.

The main challenges here are *(i)* separating individual sounds from the audio mixture and associating each soundtrack with its visual counterpart, *(ii)* estimating and tracking the spatial position of the actual sounding objects.

To address these problems, the first step is to recognize the sounding objects and their movement on the sampled video frames with the corresponding sound sources using an off-the-shelf object detection model [19] to get the object category and the 2D position: $[x, y]$ of each object in the video frame. Meanwhile, We implement a sound separation model [17, 18] to obtain the audio for each independent sound source. To match the separated soundtrack and the corresponding visual object, we apply a pre-trained audio tagging model [9] to infer the sound event category for each soundtrack, which is used for querying the sounding objects among all detected objects.

So far, the pipeline has predicted the 2D coordinate of sounding objects. To estimate the depth of each object, we use a depth estimation model [8] to estimate the depth for each pixel in a frame. Then, we use the $[x, y]$ coordinate of each object to query the depth matrix and get the corresponding $[z]$ coordinate. The complete $[x, y, z]$ coordinate of the sounding object, together with its corresponding monaural audio, will be used in the spatial audio rendering module to render the output for each audio channel in real-time.

### 3.2 Interfaces to Support User Repairs and Expressiveness

A highlight of Mimosa is that it supports human-in-the-loop error discovering and repairing of the model's predictions. This section will discuss how Mimosa (1) presents the *intermediate* results produced by each module in the pipeline in ways that amateur users can understand and validate; (2) empowers amateur users to fix any inaccuracies in the predictions of models and "go beyond the ground truth" to create their desired customized spatial audio effects.

To enable effective human-AI collaboration with imperfect ML models, an important challenge is to empower users to validate the model's *intermediate* results and make repairs when needed. To achieve this, the key is to (1) present the model's intermediate results in a form that amateur users can understand [11]; (2) provide an easy-to-use interaction method for an amateur user to express their desired result for each intermediate step to repair model inaccuracies [10].

As we can see in Figure 1, the inferred positions of sounding objects are visualized in an 2D overlay on top of the video preview (Figure 1A) and in a simulated 3D space (Figure 1B) The results of sound track separation and audio tagging are also displayed and visualized in a sound track panel (Figure 1C). The 2D overlay makes it easy for users to identify any discrepancies in sound object localization (as the dots would be misaligned or mismatched with the underlying visual objects). When there is a discrepancy, the user can simply drag the dot to match its corresponding visual object. In the meantime, the 3D simulation helps the user to easily understand the inferred positions of sounding objects relevant to the view point (the camera icon). The interaction widget provided in the 3D simulation also supports the moving and rotation of each object *or* the camera itself on individual axes, allowing finer granularity controls. The user may also directly edits the model outputs in the sound track panel (Figure 1C).

## References

1. Baldis, J.J.: Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 166–173. CHI '01, Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/365024.365092, https://doi.org/10.1145/365024.365092
2. Coleman, P., Franck, A., Jackson, P.J., Hughes, R.J., Remaggi, L., Melchior, F.: Object-based reverberation for spatial audio. Journal of the Audio Engineering Society **65**(1/2), 66–77 (2017)
3. dalton, r., tobin, j., grunzweig, d.: rondo360: dysonics' spatial audio post-production toolkit for 360 media. journal of the audio engineering society (september 2016)
4. Dede, C.: Immersive interfaces for engagement and learning. science **323**(5910), 66–69 (2009)
5. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 324–333 (2019)
6. Horvitz, E.: Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. pp. 159–166 (1999)
7. Hutchins, E.L., Hollan, J.D., Norman, D.A.: Direct manipulation interfaces. Human–computer interaction **1**(4), 311–338 (1985)
8. Kim, D., Ga, W., Ahn, P., Joo, D., Chun, S., Kim, J.: Global-local path networks for monocular depth estimation with vertical cutdepth. arXiv preprint arXiv:2201.07436 (2022)
9. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2880–2894 (2020)
10. Li, T.J.J., Chen, J., Xia, H., Mitchell, T.M., Myers, B.A.: Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. p. 1094–1107. UIST '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3379337.3415820, https://doi.org/10.1145/3379337.3415820
11. Li, T.J.J., Radensky, M., Jia, J., Singarajah, K., Mitchell, T.M., Myers, B.A.: Pumice: A multi-modal agent that learns concepts and conditionals from natural language and demonstrations. In: Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology. p. 577–589. UIST '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3332165.3347899, https://doi.org/10.1145/3332165.3347899
12. McCormack, L., Politis, A.: Sparta & compass: Real-time implementations of linear and parametric spatial audio reproduction and processing methods. In: Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio. Audio Engineering Society (2019)
13. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. Advances in neural information processing systems **31** (2018)
14. Newell, P.: Recording studio design. Routledge (2017)
15. Rumsey, F., McCormick, T.: Sound and recording: applications and theory. Routledge (2021)

16. Sánchez, J., Flores, H.: Memory enhancement through audio. In: Proceedings of the 6th international ACM SIGACCESS conference on Computers and accessibility. pp. 24–31 (2003)
17. Tian, Y., Hu, D., Xu, C.: Cyclic co-learning of sounding object visual grounding and sound separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2745–2754 (2021)
18. Wisdom, S., Erdogan, H., Ellis, D.P., Serizel, R., Turpault, N., Fonseca, E., Salamon, J., Seetharaman, P., Hershey, J.R.: What's all the fuss about free universal sound separation data? In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 186–190. IEEE (2021)
19. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
20. Zhou, H., Xu, X., Lin, D., Wang, X., Liu, Z.: Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In: European Conference on Computer Vision. pp. 52–69. Springer (2020)
21. Çamcı, A.: Some considerations on creativity support for vr audio. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 1500–1502 (2019). https://doi.org/10.1109/VR.2019.8798210