

Passing: An Endless Journey through Reconstructed Spacetime with AI-Generated Sound

Akira Takahashi , Chihiro Nagashima , Zhi Zhong , Shusuke Takahashi , Yuki Mitsufuji
Sony Group Corporation, Tokyo, Japan



Figure 1: *Passing*, a collaborative visual installation created with visual artist Ryu Furusawa, exhibited at NEORT++ in Tokyo from April 17–26, 2026 ([1]).

Abstract

This paper introduces *Passing*, an interactive audiovisual installation that generates an endless journey from a single continuous monorail-window recording by reconstructing it as a spatiotemporal volume. Rather than replaying the footage linearly, the work resamples its spatial and temporal structure along nonlinear trajectories, producing a continuously passing landscape whose depth, speed, and temporal order become unstable. A camera-based viewer-presence detection system estimates whether a viewer is present in the viewing zone and uses this presence state to influence transitions among rendered video sequences. The resulting video stream is fed into SpecMaskFoley, a real-time video-to-audio synthesis model that generates a synchronized soundscape for the reconfigured image. The model is not used to reconstruct an objectively correct soundtrack, but functions as a speculative listener, proposing a possible auditory interpretation of a world whose conventional spatial and temporal premises have been disrupted. *Passing* distributes creative agency across the artist, who defines the rules of spacetime reconstruction; the AI model, which interprets the emergent visual flow as sound; and the audience, whose embodied presence influences the audiovisual trajectory. Through this structure, the work investigates how authorship and listening may be negotiated among human intention, machine inference, and audience interpretation. Artwork page: <https://ryufurusawa.com/passing>

1 Introduction

Passing is an audiovisual artwork that generates an endless journey from a single continuous monorail-window recording by digitally reconstructing its spacetime. This process transforms the familiar experience of “passing” by deconstructing the perceptual foundations of linear time and three-dimensional space upon which it relies. This radical manipulation of the visual world raises a fundamental question: *If the premises of homogeneous space and linear time have collapsed, how can sound exist, and how does listening come into being?*

Because the reconstructed video reorganizes the temporal order, motion, and spatial relations of the original recording, it no longer has a single physically correct acoustic counterpart. The task is therefore not to recover the sound of the original scene, but to generate an auditory interpretation that remains responsive to a visually coherent yet physically impossible flow.

To investigate this task, we employed SpecMaskFoley [2], a video-to-audio (V2A) synthesis model, not to reproduce an objectively correct soundtrack, but to hear how an AI model with no body of its own may respond to reconfigured spacetime. The AI acts as a speculative listener, generating sound anew for each moment of the nonlinear video and proposing an auditory experience for a world unbound by conventional physics. Creative agency is thereby redistributed among the artist, who designs the rules of spacetime reconstruction; the AI, which interprets the emergent visual flow as sound; and the audience, whose presence or absence in the viewing zone influences the visual path.

2 Related Works

2.1 Spatiotemporal Manipulation in Media Art

The visual methodology of *Passing* belongs to a lineage of media-art practices that reorganize recorded images by altering the relation between space and time. Digital slit-scan practices, as catalogued by Golan Levin [3], and broader discussions of space-time correlations in moving-image media by Susanne Jaschko [4], provide a historical and theoretical context for treating video not as a fixed sequence of frames, but as a structure in which spatial and temporal dimensions can be exchanged, stretched, or reconfigured. Interactive works such as Alvaro Cassinelli’s *Khronos Projector* [5] and Camille Utterback’s *Liquid Time* [6] further demonstrate how the viewer’s present action can be incorporated into the temporal structure of recorded video.

A more direct artistic precedent within Ryu Furusawa’s practice is *Mid Tide #3* [7], which uses high-resolution, high-frame-rate footage of waves as source material to make altered spacetime perceptible. *Passing* extends this lineage by treating a continuous monorail-window recording as a spatiotemporal volume and generating new moving images along nonlinear cross-sectional trajectories. Unlike earlier interactive temporal manipulations constrained by real-time high-resolution rendering, *Passing* prepares transformed video clips in advance and reorganizes their playback in response to viewer presence. By coupling this presence-mediated visual sequence with real-time AI sound, *Passing* connects spatiotemporal image reconstruction, machine listening, and distributed creative agency.

2.2 Speculative Video-to-Audio Synthesis

The auditory component of *Passing* draws on recent advances in V2A synthesis. Several recent V2A methods have made significant progress in generating plausible sounds from visual or multimodal input [8–15]. Among them, MMAudio [8] demonstrates strong audiovisual synchronization and high-quality sound generation from video. SpecMaskFoley [2] also achieves synchronized V2A generation while supporting real-time inference, making it suitable for interactive installation settings.

In contrast, *Passing* uses V2A synthesis as a speculative interpretive process. The aim is not to reconstruct the correct sound of a monorail scene, but to ask what kind of sound an AI model may infer from a visual world whose temporal and spatial premises have been reconfigured. In this context, synchronization supports artistic interpretation rather than objective acoustic reconstruction. Instead of being conditioned on the original footage in its recorded order, the model receives a spatiotemporally reconfigured video stream and generates a possible auditory interpretation of the altered visual world.

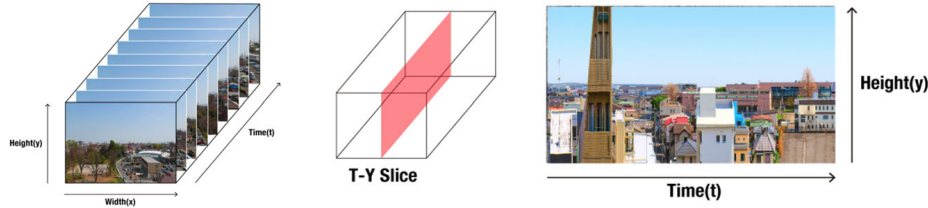


Figure 2: Spacetime reconstruction in *Passing*. Frames are stacked along the temporal axis to form a spatiotemporal volume. Ordinary playback follows this axis, whereas *Passing* generates new moving images by shifting and rotating the cross-sectional plane. As an illustrative example, a 90° rotation exchanges the temporal axis with the horizontal image dimension.

2.3 Distributed Agency in AI Art

The conceptual framework of *Passing* is related to discussions of distributed agency and authorship in AI-based creative practice. As AI systems enter artistic production, authorship is increasingly understood not as the property of a single creator, but as something negotiated among human and nonhuman actors [16]. Our previous project, the generative sound installation *Studies for* [17, 18], examined this negotiation through an AI archive trained on sound artist evala’s past works. Other works, such as Suzanne Kite’s *Listener* [19] and Justine Emard’s *Co(AI)xistence* [20], have explored AI or artificial agents as collaborators, others, or interlocutors.

Passing extends this discussion by explicitly including the audience in the agency structure. The artist defines the rules of spacetime reconstruction, the AI model interprets the resulting visual flow as sound, and the audience indirectly influences the visual path through their presence or absence in the viewing zone. The work therefore constructs a three-way agency loop in which audience behavior shapes the material that the AI interprets in real time.

3 Artwork and Interaction Design

3.1 Reconstructing Spacetime from a Continuous Recording

Passing is built from a single continuous 4K 480-fps HDR recording captured on March 27, 2025, through the window of the Tama Monorail between Tachikawa and Tama Center using a high-resolution, high-speed camera. Rather than treating this recording as a fixed sequence of frames to be played back linearly, the work treats the moving image as a spatiotemporal volume. Each frame is regarded as a photographic slice, and the sequence of frames is stacked along the temporal axis to form a virtual volume that can be traversed in ways that could not have existed in the captured scene itself. Once this volume is constructed, cross-sections can in principle be extracted at arbitrary angles. By continuously shifting and rotating the cross-sectional plane along selected trajectories, the system produces moving images that are no longer bound to the original frame order. In this sense, the transformation does not simply edit the recorded video; it resamples the temporal and spatial structure already contained within it. Fig. 2 schematically illustrates this principle using a 90° cross-sectional rotation as an example.

The recorded landscape contains multiple latent movements, including the forward movement of the train, the relative motion of nearby and distant objects, and the local motion of people, cars, and other elements within the scene. Depending on the angle of the extracted cross-section, some movements may appear to be suspended, whereas others may be stretched, reversed, or displaced. Therefore, the transformation exposes multiple temporalities latent within the original footage.

For the installation, multiple trajectory variants were generated by varying parameters such as departure timing and rotational speed, then pre-rendered and connected into a continuous playback structure. Each trajectory preserves the right-to-left flow of the passing landscape, including combinations of horizontal inversion and reverse playback. In this way, *Passing* produces an endless journey from a finite source recording by reorganizing, rather than synthetically generating, the landscape.

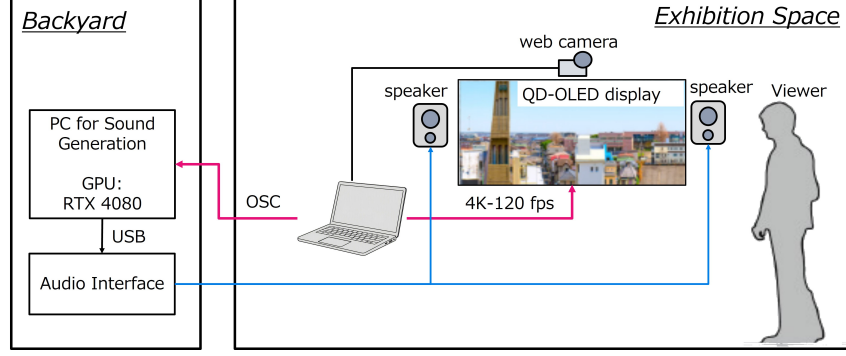


Figure 3: Installation view and system configuration of *Passing*. The work is presented on a 4K 120-fps display with stereo sound reproduction. A camera estimates viewer presence in front of the display and sends the presence state to the video playback control system, which selects rendered video clips. The selected clip determines the corresponding visual-conditioning sequence provided to SpecMaskFoley running on an RTX 4080 GPU, and the generated stereo audio is delivered through an audio interface and loudspeakers.

3.2 Audience-Mediated Sequence Reordering

In *Passing*, viewer interaction does not directly control the sound. The viewer-presence estimate is used only for video sequence control; the resulting visual sequence is then passed to the real-time sound generation model. Thus, interaction is mediated through the image: the viewer affects the visual path, and the AI model generates sound in response to the altered visual flow.

The system detects whether a viewer is present in the viewing zone and uses this state to probabilistically enable transition branches among transformed video sequences. Each sequence follows a different cross-sectional trajectory through the same spatiotemporal volume, so the work does not simply play a predetermined loop but shifts its temporal path in response to embodied presence. The rendered video content is organized into primary sequence groups and transition branches, allowing the system to sustain an endless video loop while viewer presence probabilistically enables branch transitions. The detailed group structure and sequence-transition graph are provided in Appendix B.3 and Fig. 6.

Changes in the displayed video sequence alter the temporal and spatial features available to the sound synthesis model, including motion direction, speed, event density, transition rhythm, and contradictions produced by inversion or reverse playback. Thus, the viewer affects sound only through the image, altering the world that the AI model listens to rather than operating sound as an instrument.

3.3 Installation Loop

The installation is implemented as an interactive loop connecting viewer-presence sensing, video sequence control, and real-time sound synthesis, as illustrated in Fig. 3. A camera-based viewer-presence detection system estimates whether a viewer is present in front of the display and sends only this presence state to the playback control system; no camera imagery is recorded or stored.

When a transition is triggered, the control system sends an OSC (Open Sound Control) cue to the playback engine, which selects the next clip from the rendered dataset. The selected video is displayed as 4K HDR content at 120 fps, while the corresponding visual-conditioning inputs are provided to SpecMaskFoley running on an RTX 4080 workstation. SpecMaskFoley synthesizes a synchronized stereo soundscape for the selected visual path, which is routed through an audio interface and loudspeakers. In this configuration, the viewer does not directly operate the sound-generation model. Instead, viewer presence changes the visual sequence, and the AI model generates sound in response to the resulting visual flow.

4 Real-time Sound Generation

The auditory component of *Passing* is generated in real time by SpecMaskFoley [2], a V2A synthesis model capable of producing synchronized sound for the nonlinear visual flow. Building on our previous project on continuous real-time sound generation under text/audio conditioning [18], *Passing* extends this paradigm to video-conditioned generation, where the conditioning signal is a dynamically reordered visual stream produced by reconstructed spacetime. This section summarizes the model configuration, training data, and deployment setup used in the installation.

4.1 Model Architecture and Control

SpecMaskFoley uses a pre-trained SpecMaskGIT [21] model as the audio backbone, together with a ControlNet branch and an FT-Aligner that adapts SynchFormer-based [22] temporal video features to the time-frequency ControlNet implementation. The ControlNet branch provides temporal alignment between the visual and auditory streams, allowing the model to generate sounds synchronized with both sustained visual motion and instantaneous events. A schematic overview of the architecture is provided in Appendix A.1.

For *Passing*, we omitted the CLIP visual encoder used in the original SpecMaskFoley configuration. The final system therefore relies on ControlNet for temporal video alignment and on CLAP-based audio conditioning for semantic guidance. Ambient sounds recorded inside the Tama Monorail were encoded with a pre-trained CLAP [23] audio encoder and used as conditioning information for the SpecMaskGIT backbone. This configuration supports the artistic aim of letting the audio domain guide the model’s interpretation of reconstructed spacetime, while simplifying the conditioning pathway for real-time deployment.

4.2 Training Data and Adaptation

The training data for SpecVQGAN [24], SpecMaskGIT, and Vocos [25] consist of the large-scale audio event datasets AudioSet [26] and VGGSound [27]. As a preprocessing step, we applied source separation to remove human speech components to reduce the risk of unintentionally synthesizing human-like voices.

SpecMaskFoley was developed by fine-tuning the pre-trained SpecMaskGIT model. In addition to VGGSound, the training dataset was augmented with approximately 35 min of custom audiovisual data recorded during the Tama Monorail shoot. To compensate for the limited size of this custom dataset, we applied spatial crops to the video, channel manipulations to the audio, and temporally reversed audiovisual pairs while preserving synchronization.

4.3 Real-time Deployment

Real-time generation is essential for *Passing*, because the sound must be synthesized anew for each moment of the nonlinear video loop rather than being replayed from prepared audio. For deployment, the displayed video and the visual-conditioning input were separated. The rendered clips were displayed as 4K HDR video at 120 fps, while corresponding conditioning images were precomputed and stored at 224×224 pixels and 25 fps. This deployment strategy allowed SpecMaskFoley to synthesize audio continuously for the selected visual path without extracting visual features from each displayed 4K 120-fps frame at runtime.

SpecMaskFoley was adopted because it supported continuous synchronized audio generation on the installation workstation while producing output suitable for the exhibition environment. Before the exhibition, on-site stability tests and listening checks with the artist confirmed that the A/B loop retained a stable train-interior atmosphere while the C/D transition sequences produced stronger sonic variation in response to the more distorted visual flow, guiding the final balance among synchronization, fidelity, and naturalness. In the final setup, the system generated 32-kHz stereo audio in real time and routed it through the exhibition sound system.

5 Discussion

5.1 Mediated Observation and Listening

Passing redistributes creative agency through mediation rather than direct control. The artist defines the recording, transformation rules, and constraints of the system; the viewer’s presence influences the visual path; and the AI model generates sound in response to the resulting nonlinear visual flow. Crucially, the viewer does not operate the sound as an instrument: the presence state is not mapped directly onto volume, pitch, or timbre, but instead alters the visual world that the model receives.

This mediation destabilizes both observation and listening. Because the moving image is generated by traversing a spatiotemporal volume rather than replaying a fixed sequence of frames, the landscape remains recognizable while depth, speed, viewpoint, and temporal order no longer coincide consistently. The audio is likewise not a prerecorded track attached to a stable viewpoint, but is synthesized in real time from the currently displayed visual flow. In the stable A/B sequence loop, the generated sound tended to maintain a continuous train-interior atmosphere, whereas the C/D transition sequences allowed stronger sonic variation in response to the more distorted visual flow. The artist described this effect as a sensation that the “point of observation wavers, as if in an out-of-body state.” We interpret this as a loosening of the usual correspondence between the visual point of view, the listening position, the screen, and the viewer’s body. In this sense, the AI model does not simply add sound to an image; it provides an auditory interpretation for a world in which the conditions of observation have already been destabilized.

5.2 Negotiated Soundscape and Relational Authorship

The final soundscape was developed through iterative collaboration with the artist. The aim was not to produce a realistic soundtrack or a direct sonification of visual parameters, but to hear how the AI might interpret motion, speed, and temporal instability in a world where time and space have collapsed. Adjustments to generation parameters and to the balance among fidelity, synchronization, and naturalness defined the conditions under which the model’s inference could become artistically meaningful.

The model’s limitations are central to this negotiation. Because it responds to visual motion and temporal change without possessing a body or lived spatial experience, its output may contain mismatches, ambiguities, or unexpected correspondences between image and sound. Rather than treating these as failures, *Passing* treats them as productive frictions through which the model’s interpretive agency becomes perceptible. Creative agency therefore emerges not as a property of a single actor, but as mediated, conditional, and relational.

6 Conclusion

This paper introduced *Passing*, an audiovisual artwork that generates an endless journey from a single continuous recording by digitally reconstructing its spacetime. By employing SpecMaskFoley, a real-time V2A synthesis model, we explored a fundamental question: *If the premises of homogeneous space and linear time have collapsed, how can sound exist, and how does listening come into being?* The system was designed not to reproduce a correct answer, but to hear the response of an AI model with no body of its own, positioning the model as an active participant in the creative process.

Our primary contribution lies in the investigation of distributed creative agency. We have presented a framework in which authorship is negotiated among three distinct agents: the artist, who designs the fundamental rules of the world; the AI, which proposes a real-time auditory interpretation from within that world; and the audience, whose presence influences the emergent audiovisual path and ultimately constructs personal meaning. This tripartite model moves beyond a simple human-tool relationship, revealing a dynamic and constantly negotiated process of shared imagination. *Passing* serves as a practical case study of how creative agency can be productively redistributed, rather than merely automated or diminished, through collaboration with AI.

Rather than treating real-time V2A synthesis as an isolated module, *Passing* integrates it into a continuous exhibition-scale audiovisual system in which visual reconstruction, machine inference, and audience presence mutually condition one another during live playback.

Acknowledgements

We are deeply grateful to visual artist Ryu Furusawa for giving us the opportunity to collaborate on his visual work *Passing* and to apply our sound-generation technology within his artwork. In particular, the challenge of generating sound in real time for an endless audiovisual journey through reconfigured spacetime has become an important direction that we hope to continue exploring in future work.

We also sincerely thank Hiromasa Hosotani of Sony Corporation for facilitating our introduction to visual artist Ryu Furusawa. For our research team, this collaboration began in February 2026, and Mr. Hosotani’s support and contributions leading up to the exhibition in mid-April were highly significant. We deeply appreciate these contributions.

References

- [1] Ryu Furusawa. *Passing*. <https://ryufurusawa.com/passing>, 2026. Accessed: 2026-07-28.
- [2] Zhi Zhong, Akira Takahashi, Shuyang Cui, Keisuke Toyama, Shusuke Takahashi, and Yuki Mitsufuji. SpecMaskFoley: Steering pretrained spectral masked generative transformer toward synchronized video-to-audio synthesis via ControlNet. In *WASPAA*, 2025.
- [3] Golan Levin. An informal catalogue of slit-scan video artworks and research. http://www.flong.com/archive/texts/lists/slit_scan/index.html, 2003. Accessed: 2026-07-28.
- [4] Susanne Jaschko. Space-time correlations focused in film objects and interactive video. In *Future Cinema: The Cinematic Imaginary after Film*. MIT Press, Cambridge, MA, 2003.
- [5] Alvaro Cassinelli. Khronos Projector. <https://alvarocassinelli.com/khronos-projector/>, 2004. Accessed: 2026-07-28.
- [6] Camille Utterback. Liquid Time. <https://camilleutterback.com/projects/liquid-time-series/>, 2000. Accessed: 2026-07-28.
- [7] Ryu Furusawa. Mid Tide #3. <https://ryufurusawa.com/midtide3>, 2024. Accessed: 2026-07-28.
- [8] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025.
- [9] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. *arXiv preprint arXiv:2407.05551*, 2024.
- [10] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024.
- [11] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *ICASSP*, 2025.
- [12] Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *AAAI*, 2024.
- [13] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleyrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024.
- [14] Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see—video to audio generation through text. In *NeurIPS*, 2024.

- [15] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. In *NeurIPS*, 2024.
- [16] Sofian Audry. *Art in the Age of Machine Learning*. MIT Press, 2021.
- [17] Evala. Studies for. <https://www.ntticc.or.jp/en/archive/works/studies-for/>, 2024. Accessed: 2026-07-28.
- [18] Chihiro Nagashima, Akira Takahashi, Zhi Zhong, Shusuke Takahashi, and Yuki Mitsufuji. ‘Studies for’: A human-AI co-creative sound artwork using a real-time multi-channel sound generation model. In *NeurIPS Creative AI Track 2025*, 2025.
- [19] Suzanne Kite. Listener. <https://www.kitekitekitekite.com/portfolio/listener>, 2018. Accessed: 2026-07-28.
- [20] Justine Emard. Co(AI)xistence. <https://justineemard.com/coaixistence-2/>, 2017. Accessed: 2026-07-28.
- [21] Marco Comunità, Zhi Zhong, Akira Takahashi, Shiqi Yang, Mengjie Zhao, Koichi Saito, Yukara Ikemiya, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji. SpecMaskGIT: Masked generative modeling of audio spectrograms for efficient audio synthesis and beyond. In *ISMIR*, 2024.
- [22] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. SynchFormer: Efficient synchronization from sparse cues. In *ICASSP*, 2024.
- [23] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.
- [24] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.
- [25] Hubert Siuzdak. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*, 2023.
- [26] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.
- [27] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, 2020.

Appendix A Technical Details of Real-time Video-to-Audio Generation

A.1 SpecMaskFoley Configuration and Architectural Modifications

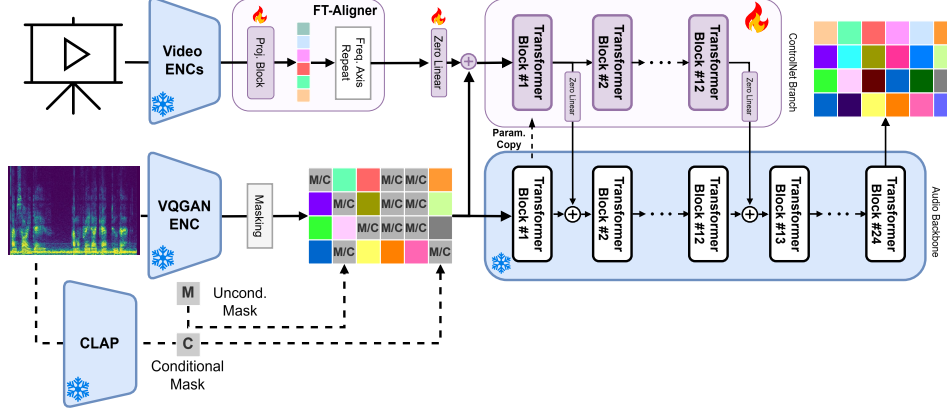


Figure 4: Overview of SpecMaskFoley. Ice icons indicate frozen modules, and fire icons indicate trainable modules. A CLAP embedding is treated as a conditional mask following SpecMaskGIT. In *Passing*, ControlNet provides temporal alignment, while CLAP-based conditioning provides semantic guidance.

This section provides additional details on the modifications made to the original SpecMaskFoley architecture [2] for its application in *Passing*. Fig. 4 shows the architecture, in which ControlNet provides temporal video alignment and CLAP-based conditioning provides semantic guidance.

While the original SpecMaskFoley model utilizes both a ControlNet for temporal synchronization and a CLIP visual encoder for semantic understanding of the video content, we deliberately omitted the CLIP visual encoder in this work for two primary reasons.

First, from an artistic standpoint, the decision was made to entrust semantic conditioning entirely to the auditory domain. Using only the CLAP audio encoder with ambient sounds recorded inside the Tama Monorail as conditioning, we anchored the generated soundscape to the stable acoustic context of a train interior. This conditioning was not intended to reconstruct the original monorail soundtrack, but to prevent the output from drifting toward unrelated or excessively artificial textures while still allowing the model to respond to the reconfigured visual motion. In this configuration, the model interprets the visual flow through a sonic rather than explicitly visual semantic context.

Second, from a technical perspective, the CLIP visual encoder proved to be a significant bottleneck for real-time inference. Its removal was crucial to ensure the stable, low-latency performance required for the interactive installation, where sound must be synthesized anew for the evolving video stream without perceptible delay. Therefore, the final architecture relies on ControlNet for temporal alignment and CLAP for semantic guidance, a configuration that satisfies both the artistic goals and technical requirements of this project.

For deployment, the video clips were converted in advance into visual-conditioning image sequences at 224×224 pixels and 25 fps. These images were cached before playback and loaded according to the selected sequence, avoiding the need to process every displayed 4K 120-fps frame during the installation.

In the *Passing* configuration, the 32 kHz SpecVQGAN tokenizer, SpecMaskGIT audio backbone, and Vocos vocoder were trained on AudioSet and VGGSound before the SpecMaskFoley adaptation, which additionally used the custom Tama Monorail recordings.

Table 1: A comparison of the model details

	SpecMaskFoley ([2])	Passing
V2A Adaptation Dataset	VGGSound 500 hours	VGGSound 500 hours + Tama Monorail recording data
Sampling rate	22.05 kHz	32 kHz
SpecVQGAN	params: 72M	params: 72M
	80 Mel bins $\times$ 848 frames	128 Mel bins $\times$ 800 frames
	265 tokens (5×53)	400 tokens (8×50)
	820 $\times$ compression	800 $\times$ compression
SpecMaskFoley	params: 300M	params: 300M
Main Network	24 Transformer blocks	24 Transformer blocks
	768 dim 8 heads attention	768 dim 8 heads attention
ControlNet	12 Transformer blocks	12 Transformer blocks
Vocoder	HiFi-GAN	Vocos

A.2 Comparison of Video-to-Audio Synthesis Performance

Fig. 5 summarizes the reported performance of SpecMaskFoley in comparison with representative V2A synthesis methods on the VGGSound test dataset. The comparison follows the evaluation in SpecMaskFoley [2], where audio synthesis quality is measured by Fréchet Audio Distance (FAD) and audiovisual temporal alignment is measured by DeSync score. Lower values are better for both metrics.

The figure is included to contextualize the choice of SpecMaskFoley as the real-time sound-generation model used in *Passing*. Unlike from-scratch V2A models, SpecMaskFoley adapts a pre-trained SpecMaskGIT audio generation backbone through a ControlNet branch and a frequency-aware temporal feature aligner. This design allows the model to achieve competitive audio quality and temporal synchronization while supporting efficient inference, making it suitable for interactive installation settings. The comparison is not intended as an evaluation of the final soundscape of *Passing*, but as background evidence for the model’s suitability as a real-time V2A component.

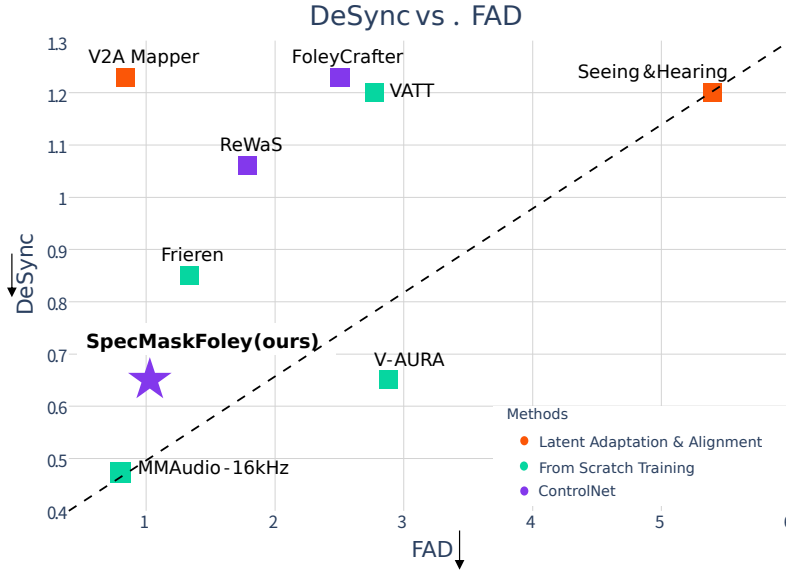


Figure 5: Audio synthesis quality and audio-video synchronization performance of representative V2A methods on the VGGSound test dataset, adapted from SpecMaskFoley [2]. FAD measures audio quality and DeSync measures temporal alignment; lower is better for both. SpecMaskFoley achieves competitive performance without from-scratch V2A training, motivating its use in *Passing*.

Appendix B Video Reconstruction and Sequence Design

B.1 Digital Coordinate Transformation

As described in Section 3.1 and illustrated in Fig. 2, *Passing* treats the recorded moving image as a spatiotemporal volume by stacking frames along the temporal axis. Ordinary playback corresponds to a cross-sectional plane advancing through this volume in the original frame order. In contrast, *Passing* changes the position and angle of this plane over time, extracting images in which spatial and temporal dimensions are partially exchanged, stretched, or reversed.

This coordinate transformation changes the perceptual structure of the image without synthesizing a new landscape. As the cross-section tilts away from the ordinary frame orientation, the simultaneity normally contained within a single frame begins to collapse, and motion traces embedded in the recording become visible. Consequently, depth, speed, viewpoint, and temporal order no longer remain consistently aligned, even though the image retains the visual continuity of recorded scenery.

This section defines the coordinate model underlying the reconstruction. The following section describes how this model was instantiated as a set of cross-sectional trajectories for the installation.

B.2 Cross-sectional Trajectory Generation

Based on the coordinate transformation described above, we generated multiple cross-sectional trajectories for the installation. Each trajectory defines how the virtual slicing plane moves through the spatiotemporal volume over time, rather than how the original frames are replayed in their recorded order.

The trajectory variants were created by varying parameters such as the departure timing and rotational speed of the cross-sectional plane. These parameters determine how different latent movements in the recorded scene become visible, including the forward movement of the train, the relative motion of near and distant objects, and the local motion of people, cars, or other elements. Depending on the trajectory, these movements may appear suspended, stretched, reversed, or displaced.

A key constraint in generating these trajectories is the preservation of the overall right-to-left flow of the landscape, including cases in which horizontal inversion is combined with reverse playback. This constraint ensures that the visual experience remains connected to the bodily perception of passing scenery, even when the internal temporal structure of the image no longer follows linear time.

The resulting trajectories were pre-rendered as video clips for use in the installation playback system. Therefore, the work does not synthesize a new landscape from scratch; rather, it reorganizes temporal and spatial structures already present in the recorded footage.

B.3 Sequence Groups and Transition Graph

The visual component of *Passing* is based on a rendered video dataset generated from a single continuous recording captured through the window of the Tama Monorail. The recorded route corresponds to the approximately 20 min trip from Tachikawa to Tama Center. The source footage was captured at 4K 480-fps HDR, while the transformed clips were pre-rendered as 4K HDR video for 120-fps playback in the installation.

The rendered clips are organized into four groups:

- **Group A:** standard forward progression (131 clips, approximately 10 s each).
- **Group B:** reverse playback combined with left–right mirroring (131 clips, approximately 10 s each).
- **Groups C and D:** branching variations from Groups A and B, consisting of clips in which temporal and spatial conditions are distorted by digital coordinate transformation. Group C transitions from forward playback toward reverse-playback and left–right-mirrored material; Group D is its inverse (Group C: 58 clips; Group D: 52 clips).

These groups define both the default playback path and the possible transition branches used during installation. Fig. 6 shows the sequence-transition structure, including the default A/B loop, presence-

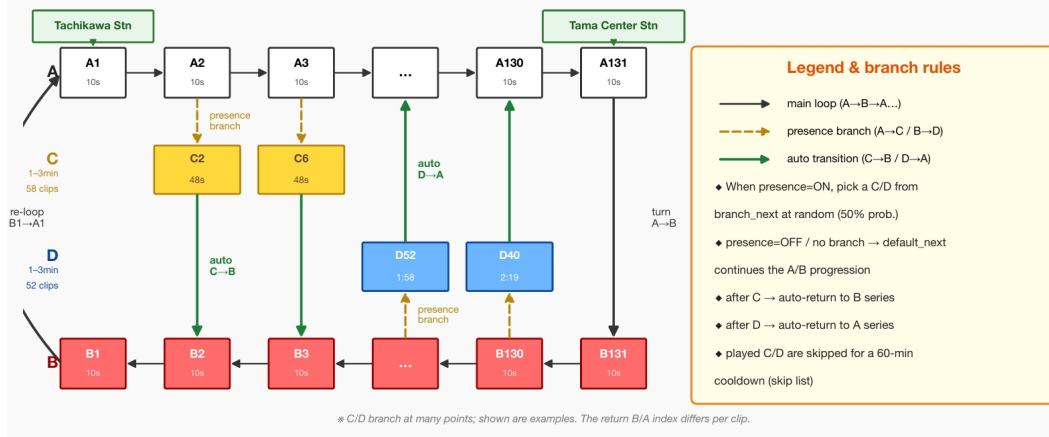


Figure 6: Transition graph of the rendered video sequences in *Passing*. Groups A and B form the default loop, while Groups C and D provide viewer-presence-enabled transition branches between the two primary states. After a C or D clip is played, playback automatically returns to the B or A series, respectively, and recently played transition clips are temporarily excluded from the candidate pool to reduce immediate repetition.

enabled C/D branches, automatic returns after transition clips, and the temporary exclusion of recently played C/D clips.

The default loop proceeds from Group A to Group B and back to Group A: A1 through A131 are played in order, followed by B131 through B1, after which playback returns to A1. This structure allows the landscape to continue flowing from right to left while the internal temporal organization alternates between forward and reversed conditions.

Viewer presence probabilistically enables the C/D transition branches. When a viewer is detected in the viewing zone, the system may select a C or D clip from the available branch candidates instead of continuing along the default A/B progression. When no branch is selected, or when no viewer presence is detected, playback continues along the default path. After a C clip is played, playback automatically returns to the B-series; after a D clip is played, it automatically returns to the A-series. To avoid immediate repetition, C/D clips that have already been played are temporarily excluded from the candidate pool for approximately 60 min.

By combining the default A/B progression with presence-enabled C/D branches, the system produces an endless video loop from a finite source recording. The same viewer staying for an extended period is unlikely to encounter the same transition repeatedly, and each viewing session emerges as a different path through the pre-rendered spatiotemporal transformations.