Seeing Sound: An Audiovisual Interpretation of Lou Reed's "Perfect Day"

Ranam Hamoud **Yuning Yao** **Derya Soydaner**
Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
{r.hamoud,y.yao.7}@umail.leidenuniv.nl
d.soydaner@liacs.leidenuniv.nl

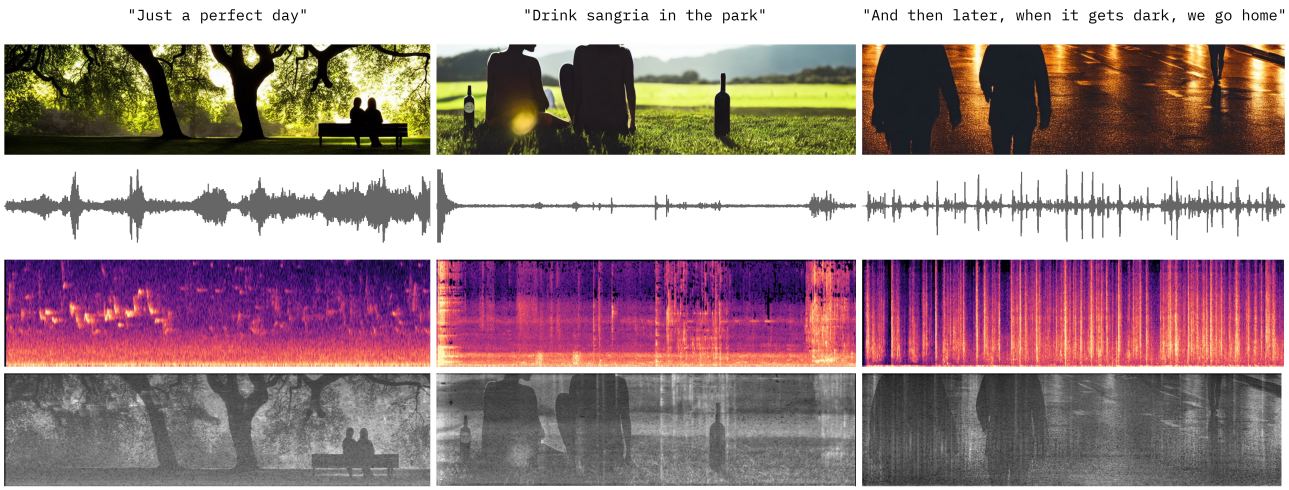


Figure 1: Example scenes from Seeing Sound. Each column corresponds to a lyric line from *Perfect Day*. From top to bottom, the rows show: the diffusion-generated image, the waveform of the field recording, its spectrogram, and the final spectrogram. In the bottom row, the generated image is visually embedded in the audio representation, producing an output that can be both seen and heard.

Abstract

Seeing Sound is an audiovisual work that treats sound as a visual and sonic medium. It asks: what does a song look like, and what does an image sound like? Interpreting Lou Reed's *Perfect Day*, the project segments the song into lyric-based scenes, each paired with a field recording and a diffusion-generated image imprinted onto its log-mel spectrogram and resynthesized into sound. The results are altered soundscapes and spectrogram-images that can be heard and read. We position the project as a study of mediated musicianship, in which lyrical ambiguity, prompting, and curation shape musical meaning alongside algorithmic generation. Grounded in everyday recordings of place and action, the system treats AI not as an autonomous creator but as part of a compositional workflow of interpretation and selection. Exploratory user feedback indicates engagement through image recognition, preserved sonic identity, and multiple narrative readings. Meaning emerges at the intersection of seeing and hearing.

1 INTRODUCTION

Can we see sound? *Seeing Sound* investigates this question through an audiovisual interpretation of Lou Reed's *Perfect Day*. We adapt a spectrogram-image pipeline to embed diffusion-generated visuals into field recordings. The outputs function simultaneously as images and as audio. Each lyric line becomes a short scene that you can look at and listen to. While generative tools make it easy to produce endless variations, this project instead asks what interpretation and curation look like when the challenge is no longer making something, but choosing what to keep. We present the work through an interactive website[1,2] in which uncovering the hidden image is part of the experience. The project treats generative models as interpretive tools whose outputs gain meaning through human recording, prompting, and selection, and not through generation alone.

The choice of source text matters. *Perfect Day* has never settled into a single reading. Some listeners hear a calm afternoon with someone close; others perceive an undercurrent of loss and loneliness. The song sustains all of these at once. Ezra Furman, writing about the album *Transformer*, describes listening to Reed as an attempt to "crack its secret code," arguing that Reed scatters "personal, emotional confessions" across the record, "largely in code" (Furman, 2018). Of *Perfect Day* specifically, Furman notes that it can feel gentle and comforting and still point somewhere darker, so that "you can't actually tell exactly what kind of song you're listening to" (Furman, 2018). This ambiguity

[1] Project website
[2] Playground

Figure 2: Selection of spectrograms from the scenes of *Seeing Sound*.

is not a problem to solve but a compositional resource. A song that resists fixed meaning gives us space to stage multiple plausible readings side by side, letting the listener move among them without arriving at a single answer (Negus and Astor, 2015; Frith, 1996). Each of our eleven scenes offers one possible interpretation without claiming to be definitive, and together they map the song's emotional range.

The project also builds on a long history of spectrogram art in popular music, where hidden images appear when a track is viewed as a time-frequency plot. In this medium, image and sound push against each other. Making a spectrogram read clearly as a picture usually introduces audible artifacts. Preserving clean audio typically weakens the visibility of the image. We treat these imperfections as part of the medium, using the tradeoff as a deliberate design constraint, not a limitation to overcome.

Lyrics provide the scene structure. For each line, we record a short field soundscape in an everyday environment and process it using an adapted version of *Images That Sound* (Chen et al., 2024). The output is a set of spectrograms that can be viewed as images and resynthesized as audio. Examples of the resulting scenes and spectrograms are shown in figs. 1 and 2. Field recording keeps the work anchored in concrete places and actions, drawing on action-sound thinking (Jensenius, 2022b), while the generative pipeline introduces visual structure that would not exist in the sound alone.

2 RELATED WORK

2.1 Spectrogram Art and the Image–Sound Tradeoff

A spectrogram represents sound energy across frequencies over time. Since it is a two-dimensional magnitude map, it can be read as an image. Musicians have exploited this for decades: Aphex Twin, Venetian Snares, and Nine Inch Nails embedded hidden pictures inside tracks that become visible in spectrogram displays (Buckle, 2022). Feaster (2023) describes these as "synesthetic sound-pictures" and identifies a recurring limitation: most outputs succeed in only one dimension, looking compelling or sounding compelling, but rarely both. This tradeoff is not only an artistic observation. Aoki et al. (2022) empirically evaluate it, assessing how reliably a visual message embedded in a spectrogram can be recovered by a viewer, and confirm that image legibility and audio quality remain in direct tension. The same representational overlap that makes spectrograms readable as images also makes them attractive to generative models. Hawthorne et al. (2022) show that diffusion models can operate directly in the spectrogram domain to synthesize high-fidelity multi-instrument audio, treating spectrogram generation as an image synthesis problem. This work shows that diffusion-based spectrogram generation can be used for audio synthesis, before text-conditioned systems later adopted similar ideas.

Audio steganography, the broader practice of hiding information in audio signals, offers a technical backdrop for understanding this space (Bender et al., 1996; Cox et al., 2007). Traditional methods aim for imperceptibility: embedded changes should be hard to hear. Our work sits on the opposite end. The embedded image is intended to be visible when the spectrogram is inspected. What matters is not concealment but the balance between visual clarity and audio quality.

2.2 Spectrograms Between Image and Audio

Classic work on signal estimation from modified short-time Fourier transforms (STFTs) made it practical to edit spectro-temporal magnitudes and reconstruct a waveform (Griffin and Lim, 1984). This turns the spectrogram into a shared canvas for image and audio generation, but also reveals the core difficulty: spectrograms do not look like everyday images, and everyday images do not sound natural when treated as spectrograms (Chen et al., 2024).

Recent diffusion-based audio systems operate on spectrogram-like representations such as log-mel spectrograms (Liu et al., 2023; Xue et al., 2024). Chen et al. (2024) build on this with a zero-shot method that jointly denoises a shared latent using an image diffusion model (Stable Diffusion) and a text-to-spectrogram model (Auffusion) in parallel. Because both models operate in the same latent space, the result is a single output that satisfies both image and audio objectives. Their method generates both modalities from text prompts. In contrast, our work replaces generated audio with an existing field recording and applies the image as a post-hoc spectrogram mask, preserving the original phase and perceptual identity of the recording.

2.3 Generative Tools and Creative Curation

Prompt-driven generative systems support a workflow of variation and selection: creators generate many outputs, then curate and arrange them. Technologies and instruments shape what feels possible to write, play, and perform (Born, 2005; Magnusson, 2019). In interactive machine learning, the algorithm can act as a creative partner whose value depends on how users steer and judge it over time (Fiebrink and Caramiaux, 2018). Authorship often lies in these human choices not in the model itself (Hertzmann, 2020). Eigenfeldt (2024) argues that artists can fully exploit generative AI systems, since the value of outputs depends on deliberate selection, contextual judgment, and the model's ability to produce variations.

Diffusion models fit naturally into this pattern. They refine noise step by step into an output (Ho et al., 2020), and latent diffusion makes this practical at high resolution (Rombach et al., 2022). Vision-language models such as CLIP (Radford et al., 2021) strengthen the link between text and image structure, making prompt writing a key mechanism to guide results. Recent work at the intersection of AI audio and artistic practice shows how these tools are adopted in compositional contexts: Yang and Ikeshiro (2024) use large audio AI models alongside field recordings and traditional Chinese instruments to generate

fixed-media accompaniment, navigating similar challenges of bridging text-conditioned AI outputs with real-world sound material. In our project, text plays a double role: it conditions image generation, but it also structures the narrative, since each prompt responds to a lyric line and a field recording.

2.4 Field Recording and Grounding

Generative systems produce many variations, but place, action, and material detail can get lost in the process. Soundscape studies and field-recording practices offer an alternative, where environmental sound is tied to place and social life (Schafer, 1994; Truax, 2008). Soundwalking describes audio capture as a hands-on practice shaped by timing, microphone placement, and attentive listening (Westerkamp, 2007). Jensenius (2022b) formalizes this through the concept of action-sound couplings, where each sound is the trace of a concrete action in the world, and demonstrates it through everyday recording practice (Jensenius, 2022a). In a generative workflow, field recordings can anchor outputs in specific places and actions and reveal how model behavior changes with real, imperfect spectral textures.

Our project brings these threads together: field recording grounds the output in place and action, lyric ambiguity provides scene-level structure (Negus and Astor, 2015; Frith, 1996), and the spectrogram serves as the shared canvas between image and sound. The following section describes the pipeline in detail.

3 METHOD

We adapt *Images That Sound* (Chen et al., 2024) to operate on externally recorded audio. Whereas the original method jointly denoises a shared latent representation to produce both modalities from text (see section 2.2), our pipeline decouples them: an image is generated independently via text-conditioned diffusion, then applied as a post-hoc magnitude mask to the spectrogram of a field recording, with the original phase preserved for resynthesis. This decoupling makes the approach applicable to any input audio signal.

For each lyric segment, the pipeline consists of three stages: *(i)* capturing a field recording, *(ii)* generating an image using a text-conditioned diffusion model, and *(iii)* imprinting the image as a mask into the recording's log-mel spectrogram, preserving its original phase. Figure 3 summarizes the full process. The following subsections detail each stage, with further elaboration on spectrogram representation, prompt design, and visualization.

3.1 Audio Capture and Preparation

We divided Lou Reed's *Perfect Day* into eleven segments aligned with lyric lines. For each segment, we captured a field recording in everyday environments (e.g., park ambience with wind and distant voices; footsteps on wet pavement at night; interior room tone with rain on windows). The recordings were captured on an iPhone 15 using the Voice Memos app (built-in microphone), in mono, and exported as lossy M4A audio. We converted the recordings to PCM WAV for editing and light cleaning, and resampled them to 16 kHz prior to analysis.

Our goal was not studio-quality audio, but sound that carries the irregularities and spatial cues of everyday settings. Following the action-sound framing discussed in section 2.4, each source recording was chosen such that its gesture, texture, or situational cue matched the emotional tone of the corresponding lyric segment. The varied spectral structures also benefit the pipeline: the model responds differently to textures shaped by movement (e.g., footsteps) than to those shaped by a static environment (e.g., room tone), contributing to visual diversity across outputs.

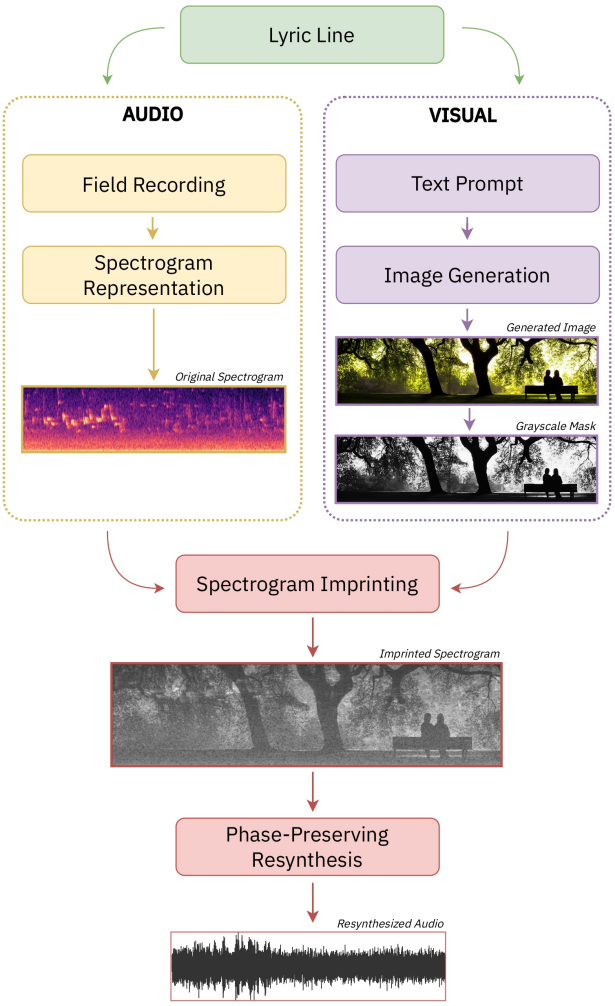


Figure 3: Overview of the pipeline. The audio branch (left) converts a field recording $x(t)$ into an STFT, separating magnitude $|X|$ and phase Φ_{orig}, and maps the magnitude to a normalized log-mel spectrogram S_{orig}. The visual branch (right) uses Stable Diffusion to generate an RGB image I, converted into an inverted grayscale mask $\tilde{I}$. This mask modulates the spectrogram element-wise (eq. (1)), embedding the image into the audio. The modified magnitude is recombined with Φ_{orig} (eq. (2)), and inverse STFT reconstructs the output waveform $y(t)$.

3.2 Spectrogram Representation

Given a waveform $x(t)$, we compute the short-time Fourier transform (STFT) $X(k, n)$ and separate the magnitude and phase as $|X(k, n)|$ and $\Phi_{\text{orig}}(k, n) = \angle X(k, n)$. We use a Fast Fourier Transform (FFT) size of 2048, a Hann window of length 1024 samples, and a hop size of 160 samples (10 ms at 16 kHz).

The magnitudes are projected to the mel scale using a 256-band mel filter bank over 0–8 kHz and log-compressed, yielding a log-mel spectrogram M.

We then apply fixed-range normalization to obtain a spectrogram canvas $S_{\text{orig}} \in [0,1]^{256\times N}$ that is compatible with image-like processing. Specifically, we linearly map log-magnitudes to $[0,1]$ using fixed lower and upper bounds, and flip the mel-frequency axis so that the representation can be treated as an image with origin at the top-left corner. The original STFT phase Φ_{orig} is retained for reconstruction.

3.3 Prompt Design

For each lyric segment, we wrote a descriptive prompt specifying scene content, figures, lighting, and the emotional tone suggested by the corresponding field recording. Prompts were refined through repeated trials: we generated batches of spectrogram-images, inspected them visually and aurally, and adjusted phrasing or added light negative prompts when early outputs contained unwanted elements. This iterative process guided the selection of images that reflect the atmosphere of each scene.

Visually, the prompting style draws on Lou Reed's own photographic work, particularly high-contrast city scenes and shadowed portraits.[3] Due to the stochastic nature of the model, the same prompt-audio pairings produced diverse outputs. The workflow therefore alternated between prompting, listening, and visual selection, using the model's variability as a space of possibilities from which the final narrative was curated.

3.4 Image Generation

For each scene, we generate an RGB image I (multiple candidates are produced per scene) using the publicly available Stable Diffusion v1.5 checkpoint[4] (Rombach et al., 2022). Text prompts are encoded with the model's CLIP text encoder (Radford et al., 2021) to produce the conditioning signal. Sampling is performed with Denoising Diffusion Implicit Models (DDIM) (Song et al., 2021), a deterministic sampling method that reverses a learned noise process in a fixed number of steps, each step subtracting a predicted portion of noise to bring the output closer to a clean image.

We run 100 such steps starting from Gaussian noise in latent space, with classifier-free guidance (Ho and Salimans, 2021) at scale $s = 10$. The resulting clean latent is decoded by the VAE decoder, which expands it by a factor of eight in each spatial dimension to produce the final image. The images are generated at a resolution of 256×1024 and resized as needed via bilinear interpolation to match the resolution of the spectrogram canvas ($256 \times N$).

3.5 Spectrogram Imprinting

To align the image and spectrogram, we convert I to grayscale (channel-mean) to obtain $I_{\text{gray}} \in [0,1]^{256\times N}$. We invert the mask so that dark regions in the original image suppress audio energy, making the image visible when the spectrogram is viewed as a picture: bright image regions leave the audio largely untouched and dark regions attenuate it. Denoting the inverted mask by $\tilde{I} = 1 - I_{\text{gray}}$, we apply the image as an element-wise modulator:

$$S_{\text{mod}} = S_{\text{orig}} \odot (1 - \lambda\, \tilde{I}), \tag{1}$$

where $\odot$ denotes element-wise multiplication and $\lambda \in [0,1]$ is a scaling factor that determines how strongly the image mask modifies the spectrogram. Pixels that are dark in the original image ($\tilde{I} \approx 1$) are attenuated most strongly, with their magnitudes scaled by approximately $1-\lambda$, while pixels that are bright ($\tilde{I} \approx 0$) leave S_{orig} nearly unchanged.

We selected λ in eq. (1) through iterative comparison across a representative subset of scenes. We tested values from 0.3 to 0.8 in increments of 0.1 and inspected each result both as a spectrogram-image and as a resynthesized waveform. Lower values, especially below 0.4, left the image difficult to distinguish from the underlying spectral texture. Higher values, especially from 0.7 upward, introduced strong attenuation across time–frequency regions, which produced audible artifacts and reduced perceptual similarity to the source recording. Values between 0.45 and 0.55 gave similar results, with only minor differences in visual clarity and audio fidelity. We therefore fixed $\lambda = 0.5$ for all eleven scenes. This value lies in the most stable part of the tested range and supports a consistent workflow across the project.

3.6 Phase-Preserving Resynthesis

To reconstruct audio, we undo the spectrogram normalization to recover a log-mel representation, and then exponentiate to recover mel magnitudes. We map these back to a linear-frequency magnitude spectrum $|X_{\text{mod}}(k,n)|$ by solving for a non-negative least-squares approximation of the inverse mel projection (McFee et al., 2015). We then construct a complex spectrum by combining the modified magnitudes with the original phase:

$$X_{\text{recon}}(k,n) = |X_{\text{mod}}(k,n)|\, e^{j\Phi_{\text{orig}}(k,n)}, \tag{2}$$

and apply the inverse STFT to obtain the waveform $y(t)$. Preserving Φ_{orig} reduces artifacts relative to iterative phase estimation (Griffin and Lim, 1984) and maintains perceptual similarity to the source recording.

3.7 Visualization

The work is presented as an interactive website.[5] We provide a playground that exposes the underlying pipeline step by step for experimentation with lyric selection, field recordings, prompting, and spectrogram imprinting.[6] Each scene begins covered by a layer of visual noise: roughly 30,000 particles arranged in a uniform grid on an HTML canvas positioned above the spectrogram-image. When the user moves the cursor, a repulsive force displaces nearby particles, and they spring back toward their original positions with easing and drag once the cursor moves away. Pushing the particles aside reveals the spectrogram-image underneath, so the picture is revealed gradually through the user's movement. This mirrors the idea of denoising, bringing hidden structures into focus, and turns listening and viewing into an active experience. The gradual reveal encourages the user to slow down and notice how sound, lyric, and image connect as the visual noise is cleared and the narrative comes together, as illustrated in fig. 4.

4 USER STUDY DESIGN

We conducted an anonymous online survey (see Appendix for full survey details) with 31 participants whose backgrounds ranged from no involvement in music or visual art,

[3] See Lou Reed's photography at Steven Kasher Gallery and Adamson Gallery.

[4] Stable Diffusion v1-5 model card

[5] Interactive viewer

[6] Playground

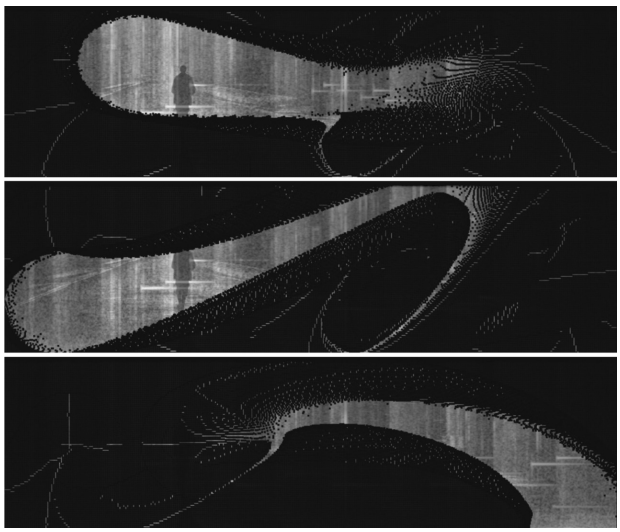
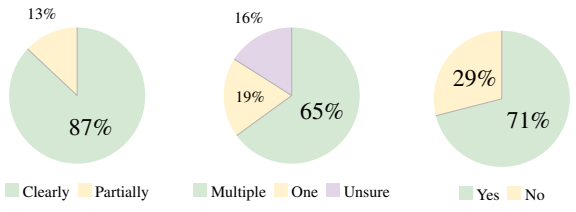


Figure 4: Example of the interactive visualization for one scene, showing three moments of the visual noise (particle layer) being cleared by user cursor movement.

to strong personal interest and formal training. The survey link was circulated online to a heterogeneous audience that included music fans, practitioners and researchers in music, sound, and visual art, as well as participants without a particular prior connection to these areas. Participants were not compensated and completed the study remotely on their own devices. They explored the interpretation of *Perfect Day* at their own pace before answering questions on image recognition, audio quality, interpretive ambiguity, and the effect of the interactive visualization. The survey combined Likert-scale ratings with optional open-ended responses to capture both quantitative and qualitative impressions.

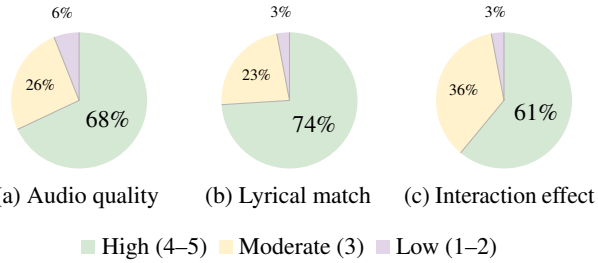
5 RESULTS AND DISCUSSION

Most participants (87%) reported that they could clearly recognize visual scenes in the spectrograms, and none reported failing to recognize them entirely (fig. 5a); the full set of categorical responses is shown in fig. 5. Ratings for audio quality (fig. 6a), lyric–scene match (fig. 6b), and the interactive visualization (fig. 6c) are also largely positive, with most responses falling in the high range (fig. 6). Together, these results suggest that the image–sound tradeoff works effectively at the chosen setting: the spectrogram-images remain visually legible, and the resynthesized audio retains enough of the source recordings to support the scene.



(a) Recognition (b) Ambiguity (c) Contrast

Figure 5: Categorical survey responses. *(a)* Whether participants can recognize the visual scenes in the spectrograms. *(b)* Whether the scenes are open to multiple interpretations. *(c)* Whether different scenes evoke noticeably different emotional impressions.



(a) Audio quality (b) Lyrical match (c) Interaction effect

Figure 6: Likert-scale responses. Most ratings fall in the moderate-to-high range.

Around two-thirds of the participants described the scenes as open to more than one reading (fig. 5b), and 71% reported that different scenes evoked noticeably different emotional impressions (fig. 5c). This matters because *Seeing Sound* does not aim to resolve *Perfect Day* into a single visual storyline. Instead, it stages a set of lyric-based scenes that preserve the song's inherent lyrical ambiguity and allow multiple interpretations to coexist across the sequence.

The open-ended responses reinforce this pattern. Responses to the open-ended questions about standout scenes and contrasting impressions are summarized using word-frequency processing with light normalization and synonym merging. Figure 7 shows the 25 most frequent terms by total mentions across responses. Participants used a wide range of emotional and descriptive terms when reflecting on scenes that stood out and contrasts across the work. These responses suggest that the scenes invite distinct interpretations. Participants did not only report recognizing images in the spectrograms. They also attached different moods, situations, and narrative possibilities to different parts of the sequence. In this sense, the work produces interpretive variation alongside visual variation.

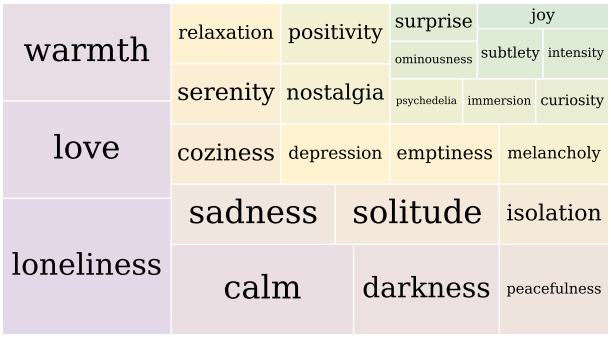


Figure 7: Treemap of emotional and descriptive terms from open-ended responses about standout scenes and contrasting impressions across scenes. Rectangle area reflects frequency.

A recurring contrast emerged between the warmth of the earlier scenes and the darker undertow of the later ones. This pattern aligns with readings of *Perfect Day* as a song that can feel gentle and comforting but also point somewhere darker (Furman, 2018). Several participants explicitly noted that the work changed how they understood the song, shifting it from a straightforward pleasant narrative to a more layered and ambivalent reading. The project therefore does not reduce ambiguity. It makes that ambigu-

ity perceptible through an audiovisual sequence of curated scenes.

Participants also responded positively to the interactive visualization (fig. 6c). Their comments point to three connected effects. First, several participants reported that the gradual reveal drew their attention more closely to visual and sonic detail. Second, comments suggested that the interaction framed the experience as one of uncovering each scene rather than receiving it all at once. Third, participants described a clearer sense of pacing, since each scene emerged progressively through their own movement. The visualization therefore does more than display the outputs. This suggests that meaning emerges through the intersection of lyric, sound, image, and viewer.

When asked whether the work changed how they hear *Perfect Day*, many participants described deeper engagement with the lyrics, stronger awareness of contrast across scenes, and a sense of moving through the song as an active listener. These responses support our framing of the project as mediated musicianship (Eigenfeldt, 2024; Hertzmann, 2020). The contribution lies not in treating AI as an autonomous creator, but in using it within a compositional workflow shaped by field recording, prompting, interpretation, selection, sequencing, and presentation. The results therefore support our central claim: the contribution of *Seeing Sound* lies in how generative outputs are curated into an audiovisual interpretation that can be both seen and heard.

6 LIMITATIONS

Our pipeline reveals a clear image–sound tradeoff. Stronger magnitude attenuation improves spectrogram-image clarity but makes the audio sound less like the original recording. Although preserving the original phase helps retain the recording's identity, large magnitude modifications can introduce phase inconsistency and audible noise. The mel-spectrogram inversion introduces additional reconstruction error, making it difficult to separate imprinting artifacts from inversion artifacts.

The method is also sensitive to design and recording choices. Changes in prompts, random seeds, or guidance settings can alter the generated image, and microphone placement, background noise, and spectral density influence how legibly the image appears in the spectrogram. The playground suggests that the pipeline can be applied beyond a single song, but we have not yet systematically evaluated its generalizability across different settings.

Our outputs reflect this tradeoff. At our chosen value of the scaling factor, the modification is moderate, but some scenes still change the spectral balance of the recording because dark image regions suppress perceptually important frequency bands. Tuning of this factor scene by scene could improve this balance, but it would reduce reproducibility. In addition, mel-based resynthesis is inherently lossy, since reconstruction cannot recover spectral detail discarded by the projection.

Our evaluation is primarily qualitative. We selected outputs through iterative listening and inspection without formal listening studies or objective audio–quality metrics, and therefore we do not make broad claims about audience perception. The scope is deliberately narrow: one song, eleven scenes, smartphone recordings at 16 kHz, and Stable Diffusion v1.5. Prompt design and curation are inherently subjective, and different authors would likely produce different results. We present the method itself as the project's main contribution, with the outputs as examples of its use.

7 CONCLUSION

Seeing Sound treats the spectrogram as a shared canvas where image and audio coexist. It positions generative AI as part of a compositional workflow shaped by lyrical ambiguity, field recording, prompt design, and curation. Meaning emerges less from what the model produces and more from the human decisions that surround it: which recording to make, which prompt to write, which output to select, and how to present it. The interactive element extends this logic to the audience, turning listeners into active interpreters who uncover each scene at their own pace and construct their own reading of a song that has always resisted a single one.

Future work could extend the approach to other source texts and musical forms, explore adaptive control of the imprinting strength, and investigate higher-fidelity resynthesis methods. Larger listener studies would help clarify how visual embedding shapes auditory perception across different audiences and levels of musical expertise. *Seeing Sound* opens a space for AI in creative practice shaped by human interpretation, curation, and decision-making.

8 ETHICS STATEMENT

Before beginning the survey, each participant was shown a consent form explaining what data would be collected, how it would be used, and their right to withdraw. Participation was voluntary, and participants could stop at any time prior to submission. No names, email addresses, or other directly identifying information were collected. Consent was explicitly recorded, with each participant required to select a consent option before proceeding. All responses are stored anonymously and are reported only in aggregated or anonymized form. We used publicly available pretrained models for image generation and did not train or fine-tune any AI models on participant data, self-recorded audio, or third-party audio material. Any short third-party excerpts included in the outputs were limited to less than ten seconds and used only as surrounding texture or to support the field recording in the artistic output; they were not used as training data for any AI model. The full survey is presented in the Appendix.

REFERENCES

Aoki, N., Ikeda, K., Yasuda, H., Shen, Y., and Hou, J. (2022). Some evaluations on spectrogram art communications exchanging secret visual messages. In *Proceedings of the 6th International Conference on Digital Signal Processing (ICDSP)*, pages 73–77.

Bender, W., Gruhl, D., Morimoto, N., and Lu, A. (1996). Techniques for data hiding. *IBM Systems Journal*, 35(3-4):313–336.

Born, G. (2005). On musical mediation: Ontology, technology and creativity. *Twentieth-Century Music*, 2(1):7–36.

Buckle, B. (2022). Spectrogram art: A short history of musicians hiding visuals inside their tracks. https://mixmag.net/feature/spectrogram-art-music-aphex-twin.

Chen, Z., Geng, D., and Owens, A. (2024). Images that sound: Composing images and sounds on a single canvas. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 85045–85073.

Cox, I. J., Miller, M. L., Bloom, J. A., Fridrich, J., and Kalker, T. (2007). *Digital Watermarking and Steganography*. Morgan Kaufmann, second edition.

Eigenfeldt, A. (2024). The Artist in A.I. Art. In *Proceedings of the International Conference on AI and Musical Creativity (AIMC)*.

Feaster, P. (2023). Creating synesthetic sound-pictures with generative AI. https://griffonagedotcom.wordpress.com/2023/12/23/creating-synesthetic-sound-pictures-with-generative-ai/.

Fiebrink, R. A. and Caramiaux, B. (2018). The machine learning algorithm as creative musical tool. In Dean, R. T. and McLean, A., editors, *The Oxford Handbook of Algorithmic Music*. Oxford University Press.

Frith, S. (1996). *Performing Rites: On the Value of Popular Music*. Harvard University Press.

Furman, E. (2018). *Lou Reed's Transformer*. Bloomsbury Publishing.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

Hawthorne, C., Simon, I., Roberts, A., Zeghidour, N., Gardner, J., Manilow, E., and Engel, J. (2022). Multi-instrument music synthesis with spectrogram diffusion. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

Hertzmann, A. (2020). Computers do not make art, people do. *Commun. ACM*, 63(5):45–48.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Ho, J. and Salimans, T. (2021). Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Jensenius, A. R. (2022a). 2022, a year of sound actions. https://www.arj.no/2022/01/01/a-year-of-sound-actions/.

Jensenius, A. R. (2022b). *Sound Actions: Conceptualizing Musical Instruments*. The MIT Press.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. (2023). AudioLDM: Text-to-audio generation with latent diffusion models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR.

Magnusson, T. (2019). *Sonic writing : Technologies of Material, Symbolic and Signal Inscriptions*. Bloomsbury Academic, first edition.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. *SciPy 2015*.

Negus, K. and Astor, P. (2015). Songwriters and song lyrics: architecture, ambiguity and repetition. *Popular Music*, 34(2):226–244.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685. IEEE Computer Society.

Schafer, R. M. (1994). *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books.

Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*.

Truax, B. (2008). Soundscape composition as global music: Electroacoustic music as soundscape. *Organised Sound*, 13(2):103–109.

Westerkamp, H. (2007). Soundwalking. In Carlyle, A., editor, *Autumn Leaves: Sound and the Environment in Artistic Practice*, page 49. Double Entendre.

Xue, J., Deng, Y., Gao, Y., and Li, Y. (2024). Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4700–4712.

Yang, H. and Ikeshiro, R. (2024). Large audio AI models for fixed-media electronics in "prelude: To listening". In *Proceedings of the International Conference on AI and Musical Creativity (AIMC)*.

A ONLINE SURVEY

The survey was hosted on Qualtrics and distributed via an anonymous link. Participants could withdraw at any time before submission. After providing consent, they were asked to explore the interactive website at https://perfectday.blog/ (shown in fig. 8) at their own pace before answering the survey questions listed in table 1.

Table 1: Survey questions in the order presented to participants.

#	Question	Response format
About you		
Q1	How old are you?	Under 18 / 18–24 / 25–34 / 35–44 / 45–54 / 55–64 / 65 or older
Q2	How do you describe yourself?	Female / Male / Non-binary / Third gender / Prefer not to say / Prefer to self-describe
Q3	What is your home country?	Drop-down list of countries
Q4	What is the highest level of education you have completed?	Secondary / Some university but no degree / Bachelor's degree / Master's degree / Doctoral or professional degree
Q5	Which best describes your background?	I have formal training or professional experience in music, sound, or visual art / I have a strong personal interest in music, sound, or visual art (but no formal training) / None of the above
Image & sound		
Q6	Have you been able to recognise visual scenes (figures, places, objects) in the spectrograms?	Yes, clearly in most scenes / Partially – in some scenes but not others / No, I could not make out the images
Q7	How would you rate the overall audio quality of the scenes?	5-point Likert scale (1 = Very degraded … 5 = Clear / high quality)
Interpretation		
Q8	How well do the visual and audio scenes match the lyric lines they are paired with?	5-point Likert scale (1 = Not at all … 5 = Very well)
Q9	Which scene stood out to you the most? (You can identify it by its lyric line or number from the list above.) What mood, situation, or story did it suggest?	Open response
Q10	Did any other scene give you a noticeably different emotional impression?	Yes / No
Q11	If yes, which scene, and what did it suggest?	Open response
Q12	Thinking about the scene you described in the previous questions, did it feel as though it had one clear meaning, or could it be read in more than one way?	It suggested one clear mood or story / It felt open to more than one reading / I'm not sure
Interaction		
Q13	The images are initially hidden behind a layer of particles or visual noise that the viewer pushes aside. How much did this interactive reveal affect your experience?	5-point Likert scale (1 = Not at all … 5 = Significantly)
Q14	In what way, if any, did uncovering the image change how you engaged with the scene?	Open response
Overall		
Q15	Did this work change or add to how you hear or think about *Perfect Day*?	Open response
Q16	Any other comments or observations? (optional)	Open response

B SCENE DETAILS

Table 2 lists the lyric line, image-generation prompt, and source audio for each of the eleven scenes in *Seeing Sound*. Where applicable, short third-party excerpts were used as surrounding texture or to support the field recording. The prompts shown are the final versions used to generate the selected outputs.

Table 2: Scene-by-scene lyric lines, image-generation prompts, and source audio.

#	Lyric	Prompt	Source audio
Part One: Two			
1	Just a perfect day	Two silhouettes sitting on a park bench under large trees, soft morning light through branches, balanced composition, calm mood	Park ambience, birds chirping, gentle wind through trees
2	Drink sangria in the park	Two silhouettes sitting on grass, one pouring wine into a glass, sunlight glowing through the bottle, diagonal shadows, gentle relaxed posture	Pouring liquid into glass, glass clink, people chatting in distance
3	And then later, when it gets dark, we go home	Two silhouettes walking side by side under streetlights, reflections on wet pavement, long perspective, quiet evening atmosphere	Rain on pavement, footsteps on wet ground, distant car passing
4	Feed animals in the zoo	Two silhouettes standing behind glass, blurred animal shapes in background, soft daylight, clear geometric framing	Children laughing, animal sounds in distance, zoo ambience, echo of glass
5	Then later, a movie too, and then home	Two silhouettes sitting in front of a glowing cinema screen, subtle light outlines their profiles, triangular composition, still moment	Cinema projector sound, with a short excerpt from *Faraway, So Close!* in the background (Lou Reed cameo scene)
6	Oh it's such a perfect day, I'm glad I spent it with you	Two silhouettes sitting at a dinner table, candle between them, soft reflections on table surface, symmetrical framing, intimate atmosphere	Clinking wine glasses, quiet restaurant ambience, soft breathing, candle crackle
Part Two: One			
7	Just a perfect day, problems all left alone	One silhouette sitting alone on a park bench under a cloudy sky, large empty space around, flat light, quiet isolation	Orchestra tuning—the story recalibrates
8	Weekenders on our own, it's such fun	One silhouette standing still among moving shadows of a dancing crowd, blurred motion lights, central framing, contrast between motion and stillness	Club bass muffled from distance, crowd laughter, door closing, one footstep leaving, with a short excerpt from Niall Horan's "This Town" (Tiësto Remix)
9	You make me forget myself, I thought I was someone else, someone good	Single silhouette reflected in a rainy window, droplets distorting the shape, faint city light behind, fragmented reflection	Rain against window, light wind, faint city hum, breath near microphone
10	You just keep me hangin on	One silhouette sitting at a dinner table, empty chair opposite, candle burned out, thin beam of light across frame, strong geometry	Empty room ambience, glass on table, chair dragging on floor
11	You're going to reap just what you sow	One silhouette walking alone down a long empty street at dawn, stretched shadow forward, centered composition, quiet atmosphere	Early morning ambience, slow footsteps fading away, with a short excerpt from Fairuz's "Adesh Kan Fi Nas" (instrumental)

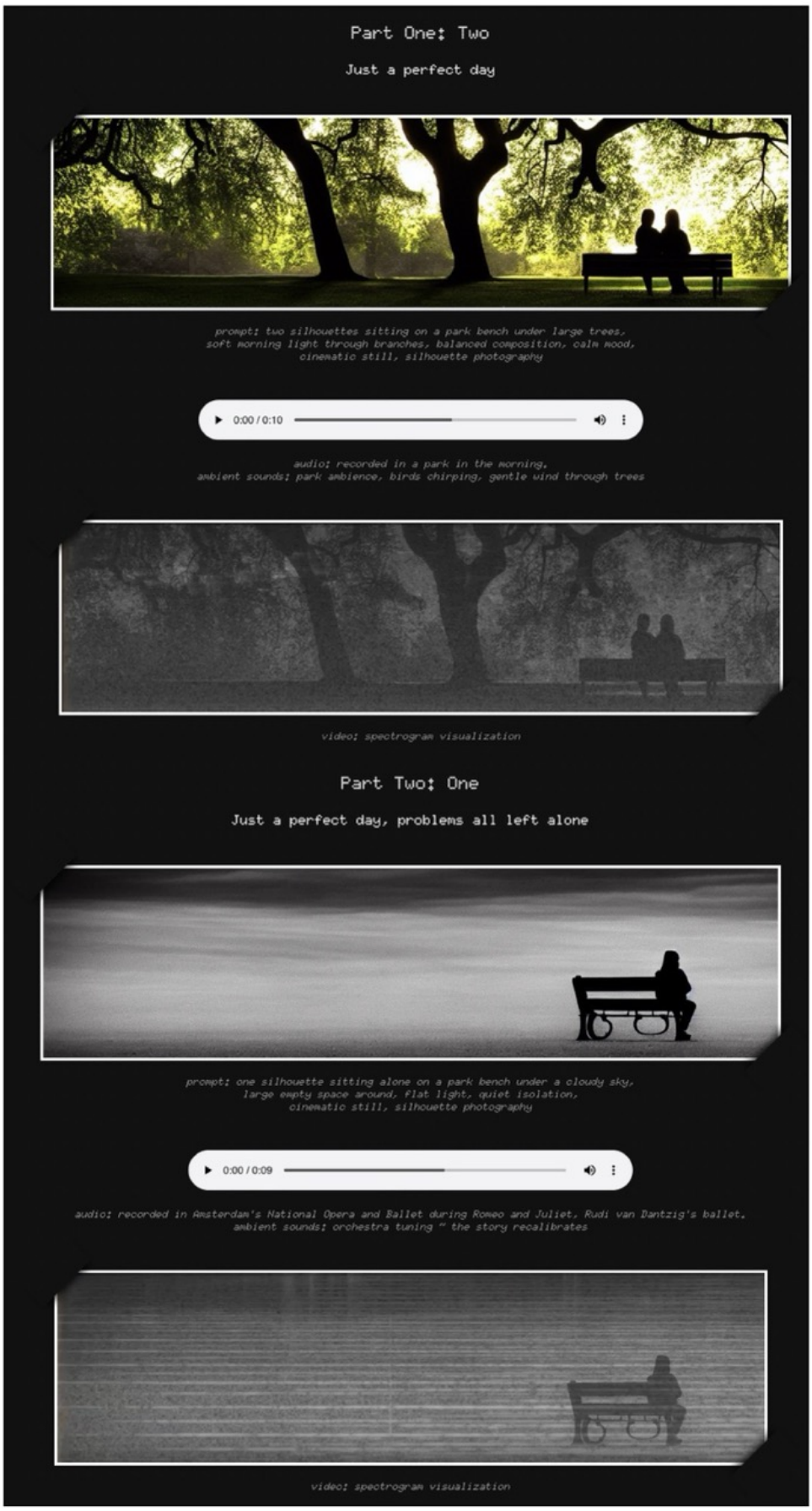


Figure 8: Album view of the website. For each scene, participants could compare the original diffusion-generated image, the original field recording, the resulting spectrogram-imprinted image, and the resynthesized audio.