Highlights in Art and Design
ISSN: 2957-8787 | Vol. 15, No. 2, 2026

Sound Trace Flow Composition: A Method for Generating Abstract Visual Forms from Acoustic Features

Yutong Mei, Zimo Dong *

Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai, China

*** Corresponding author:** Zimo Dong

Abstract: This paper introduces Sound Trace Flow Composition, a rule-based method for converting birdsong features into abstract visual structures. We extract and normalise four features: RMS energy, spectral centroid, frequency-band energy ratios, and zero-crossing rate, to influence flow lines, density, geometric fragmentation, continuity, and negative space. The composition is organised using a Bézier-based framework, and controlled randomness is incorporated to vary the composition within constraints derived from the source. The method preserves the temporal and spectral differences of different birdsong samples in a non-conventional manner by applying a shared visual grammar to them, avoiding reduction to conventional data plots. It bridges explicit sound-to-visual mappings and generative art by providing traceable relationships, constrained variation, and reproducible visual construction.

Keywords: Sound-to-Visual Mapping; Visual Abstraction; Generative Art.

1. Introduction

Sound is a constant unfolding in time, with variations in energy, frequency distribution, and transient structure. Digital signal processing allows these features to be converted into measurable values. They can be used by procedural graphics not only as data for analysis, but also as criteria for visual synthesis.

Traditional sound visualisation conveys acoustic information in the form of waveforms, spectrograms and frequency graphs. Conversely, recent cross-modal systems create recognisable visual content from audio, but the relationship between auditory elements and visual structures may become difficult to detect. This creates an unchartered region between analytical visualisation and cross-modal generation: a method that maintains a traceable connection to the original sound but allows the image to evolve as an independent artwork.

This study provides Sound Trace Flow Composition, a rule-based method to convert bird sound characteristics to picture production parameters, on the basis of ‘acoustic characteristics → generation parameters → visual structure’. These parameters regulate layered structures and explicit mapping rules ensure traceability. Bounded random variation allows compositional diversity without breaching limitations. The method attempts to satisfy three constraints: traceability to the original sound, reproducibility from identical inputs, and variety amongst generated visual pieces.

2. Literature Review

These methodologies have increased the variety of visual material and styles that sound can affect, but, the impacts of individual sound elements on specific visual structures are frequently not easily matched and instead, depend on internal latent representations. Lima et al. surveyed music visualization by input features, musical attributes, visualization strategies, interaction, and evaluation [1]. Foote showed that audio self-similarity can translate temporal repetition and structure into two-dimensional spatial relations [2]. Dzwonczyk et al. distinguished functional visualization from aesthetic visualization, which focuses on artistic creation [3]. This study follows that aesthetic direction but keeps explicit links between acoustic descriptors and visual structure. Work in music information retrieval also provides a technical basis for this kind of mapping. Tzanetakis and Cook used spectral centroid, zero-crossing rate, and energy-related features to describe temporal and spectral properties of audio [4]. Research on crossmodal correspondence has examined stable links between auditory and visual dimensions. Grill and Flexer mapped perceptual sound qualities to visual properties such as brightness, hue, regularity, and contour [5]. Spence reviewed recurring correspondences between pitch, intensity, size, brightness, and spatial position [6]. Rodrigues et al. classified auditory-visual mappings as 1-1, 1-N/N1, and N-M relationships [7]. Because the present method lets several acoustic features regulate several visual variables, it is closer to an N-M mapping.

Since the present method allows multiple acoustic features to jointly regulate multiple visual variables, its structure is closer to an N-M mapping. Recent work has increasingly employed learned cross-modal representations. Shim, Kim, and Kim used birdsong to generate species-related visual imagery through an audio encoder and generative model [8]. Sound2Scene aligns acoustic representations with a visual latent space for scene generation [9], while Lee et al. explored music-to-visual style transfer through shared semantic relations between music and paintings [10]. All these methods demonstrate the generative potential of sound, but most rely on latent representations learned by models, making the relationship between specific acoustic features and visual structures difficult to trace. Dzwonczyk et al. used RMS, spectral centroid, and spectral flux to control a diffusion model [3]; Chen, Geng, and Owens developed a shared representation in which the same two-dimensional structure can be interpreted as both image and sound [11]; Routray et al. proposed a workflow from sound-feature extraction to visual synthesis [12]; and Patra et al. examined GAN-, VAE-, and Transformer-based approaches to music-to-visual generation [13].In contrast, Liu et al.'s ChladniSonify establishes a reproducible physical correspondence between visual patterns and sound frequencies, that providing a reference for interpretable mapping [14].

Existing sound-to-vision methods can be grouped into direct visualization, explicit parameter mapping, and latent-space generation, prioritizing readability, controllability, and generative freedom, respectively. A gap remains between traceable acoustic features and integrated visual creation. Sound Trace Flow Composition addresses this gap through rule-based procedural generation. It uses RMS, spectral centroid, frequency-band energy ratio, and zero-crossing rate to control streamlines, geometric fragments, density, continuity, and negative space.

3. Method

3.1. System Overview

Sound Trace Flow Composition turns the temporal, spectral, and energy characteristics of birdsong into layered visual structures. The workflow covers audio loading, feature extraction and processing, visual segmentation, procedural generation, preview, and high-resolution rendering. The system extracts RMS energy, spectral centroid, frequency-band energy ratios, and zero-crossing rate, then smooths, clips, and normalizes them. A Bezier-based flow-line skeleton organizes the composition, and these features jointly control secondary lines, geometric fragments, textures, density, and negative space. When the analysis results, parameters, and random seed are fixed, the system reproduces the same high-resolution output.

Fig 1. Overall Workflow of the Sound Trace Flow Composition System

3.2. Audio Preprocessing and Feature Extraction

3.2.1. Audio Loading and Mono Conversion

For an input with C channels, let y_c[n] denote the sample at index n in channel c. The mono signal is defined as:

$$y_{mono}[n]=\frac{1}{C}\sum_{c=1}^{C} y_{c[n]} \quad (1)$$

The conversion keeps the original sampling rate and all the following analysis is performed on the resulting mono signal. Empty files, improper sample rates, quiet recordings and undecodable inputs are rejected before feature extraction.

3.2.2. Framing and Spectral Analysis

Let the sample rate be fs. From an analysis window of roughly 46 ms a first estimate of the desired frame length is calculated:

$$L_{\{target\}}= f_s \times 0.046 \quad (2)$$

The actual frame length L is rounded up to the next power of two within the range of the program. The hop length is defined as:

$$H = 0.25L \quad (3)$$

Thus H = 0.25L corresponds to 75% frame overlap. A short-time Fourier transform is then applied to the mono signal:

$$X_{t(f)}=\mathrm{STFT}\left(y_{mono}\right) \quad (4)$$

The magnitude and power spectra are defined as:

$$M_{t(f)}= \left|X_{t(f)}\right|,\quad P_{t(f)}= \left|X_{t(f)}\right|^2 \quad (5)$$

These spectral quantities are used to calculate the spectral centroid and the energy ratios of the designated frequency bands.

3.2.3. Acoustic Feature Definition

The system extracts frame-level RMS energy, spectral centroid, low-, mid-, and high-frequency energy ratios, and zero-crossing rate.

RMS describes local changes in signal energy:

$$\mathrm{RMS}(t)=\sqrt{\frac{1}{N}\sum_{n=1}^{N} x_t\,[n]^2} \quad (6)$$

The spectral centroid describes the overall position of spectral energy along the frequency axis:

$$C(t)=\frac{\sum_f f\cdot M_t(f)}{\sum_f M_t\,(f)} \quad (7)$$

Let the total spectral energy be:

$$E_{total}(t)=\sum_f P_t\,(f) \quad (8)$$

The energy ratio for an arbitrary frequency band B is then:

$$R_B(t)=\frac{\sum_{f\in B} P_t\,(f)}{E_{total}(t)} \quad (9)$$

For visual mapping, the system uses three working bands: 20–300 Hz (low), 300–2000 Hz (mid), and 2000 Hz to the Nyquist frequency (high). The upper limit of the high band is clipped automatically according to the sampling rate. These ranges are internal control bands for the generative system and are not proposed as a bioacoustic classification of birdsong.

Zero-crossing rate describes how frequently the signal changes sign:

$$\mathrm{ZCR}(t)=\frac{1}{N-1}\sum_{n=1}^{N-1} I\,[x_t[n]\cdot x_t[n+1]<0] \quad (10)$$

These features serve as structural control variables in Sound Trace Flow Composition and are not plotted directly as the final image.

3.2.4. Feature Post-processing

To reduce the influence of outliers and short-lived fluctuations, feature post-processing consists of percentile clipping, Gaussian smoothing, and normalization.

Extreme values are first suppressed by percentile clipping, after which each continuous feature series is smoothed with a one-dimensional Gaussian kernel:

$$\tilde{x}(t)=G_\sigma * x(t) \quad (11)$$

Then the smoothed feature is normalised to [0,1] using percentile bounds as:

$$x_{norm}(t)=\frac{\tilde{x}(t)-p_{low}}{p_{high}-p_{low}} \quad (12)$$

where here (p_{low}) and (p_{high}) are the lower and upper percentile boundaries used for normalisation.The processed series is the primary input for visual production, while the raw data are maintained for activities requiring finer local variation.

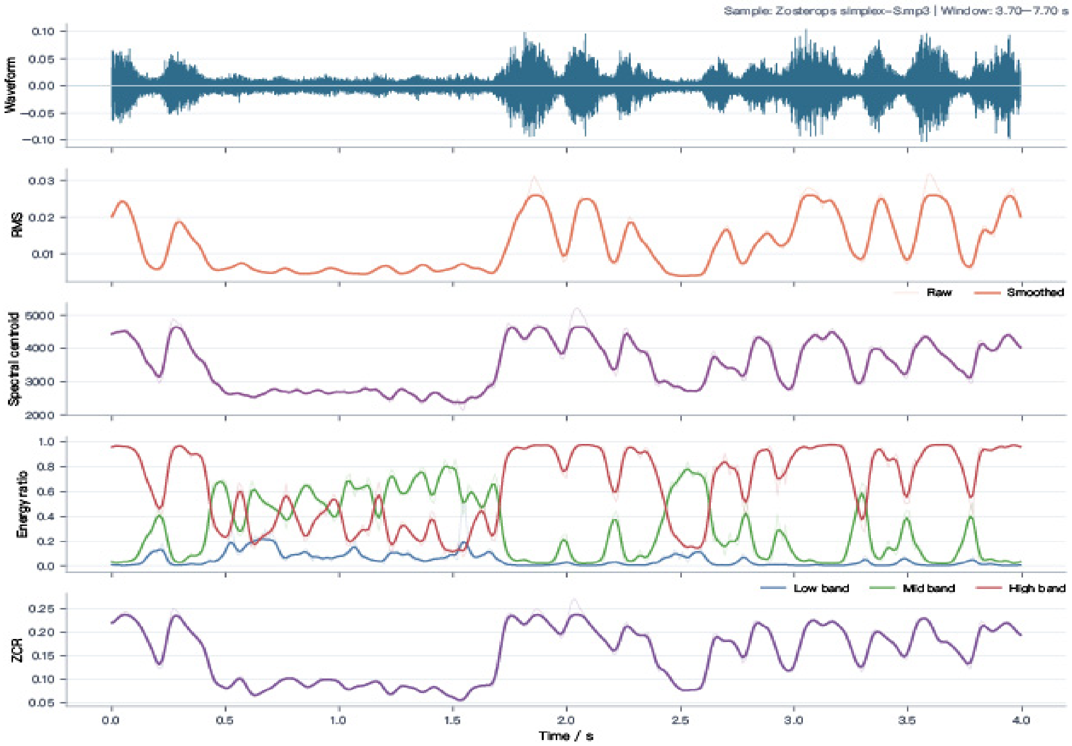

Fig 2. Bird-call acoustic feature analysis

3.3. Sound Trace Flow Composition Generation

3.3.1. Design Rationale

Sound Trace Flow Composition transforms time-varying acoustic features into a continuous visual field. It follows the hierarchy of acoustic feature → generative parameter → visual structure, with multiple features jointly controlling scale, density, direction, continuity, and local complexity.

3.3.2. Visual Segmentation and Flow-Line Skeleton

The continuous feature series is divided into visual segments to avoid mechanical frame-by-frame continuity. Each segment stores local acoustic statistics, which guide flow-line generation. The primary flow line is represented by a cubic Bézier curve:

$$P(t)=(1-t)^3P_0+3(1-t)^2tP_1+3(1-t)t^2P_2+t^3P_3, \quad t\in[0,1] \quad (13)$$

(P_0) and (P_3) are the start and end points, while (P_1) and (P_2) are control points that determine the bending direction of the cubic Bézier curve. The primary flow line establishes the large-scale direction and motion of the image. Secondary lines are generated around it and may converge, diverge, intersect, offset, or disappear locally.

3.3.3. Mapping Acoustic Features to Visual Structure

(P_0) and (P_3) define the endpoints, while (P_1) and (P_2) control the Bézier curve's direction. The primary flow line sets the overall movement, and secondary lines create local variations.

RMS controls the line weight, fragment scale and density. Control flow length, curvature, secondary lines, fine detail and direction variations of low, mid and high-frequency energy. Spectral centroid impacts vertical displacement: zero-crossing rate contributes to the enhancement of broken lines and short strokes. Low energy areas generate more negative space and broken routes. These are design rules unique to the systems, not general acoustic-visual equivalences. Detailed mappings are given in Table 1.

Table 1. Mapping of Acoustic Features to Visual Control Parameters

Acoustic feature	Visual control parameters	Generative function
RMS energy	Line weight; fragment scale; local element density	Regulates the local visual intensity and structural prominence
Low-frequency energy ratio	Primary flow-line length; flow-band width; large-scale curvature	Controls the large-scale trajectory and spatial extent of the flow structure
Mid-frequency energy ratio	Number of secondary flow lines; structural density; local connections	Regulates the complexity and connectivity of secondary linear structures
High-frequency energy ratio	Fine-line density; angular fragments; local directional variation	Increases local fragmentation, sharpness, and directional complexity
Spectral centroid	Vertical displacement of flow lines	Controls the vertical position and spatial offset of the flow structure
Zero-crossing rate	Probability of broken lines; short strokes; incomplete geometric forms	Regulates local discontinuity and fragmented visual texture
Low-energy regions	Reduced element density; path interruption; negative space	Produces sparse regions and negative space within the overall composition

3.3.4. Geometric Fragments and Linear Textures

Once the primary and secondary flow lines are built, the system samples points and tangent directions along them, to produce orientated lines, geometric fragments and textures. The structure is marked out by thick lines, flow bands are made by finer lines, and the intricacy is added by fragments and micro-strokes to produce texture, all within the general linear flow.

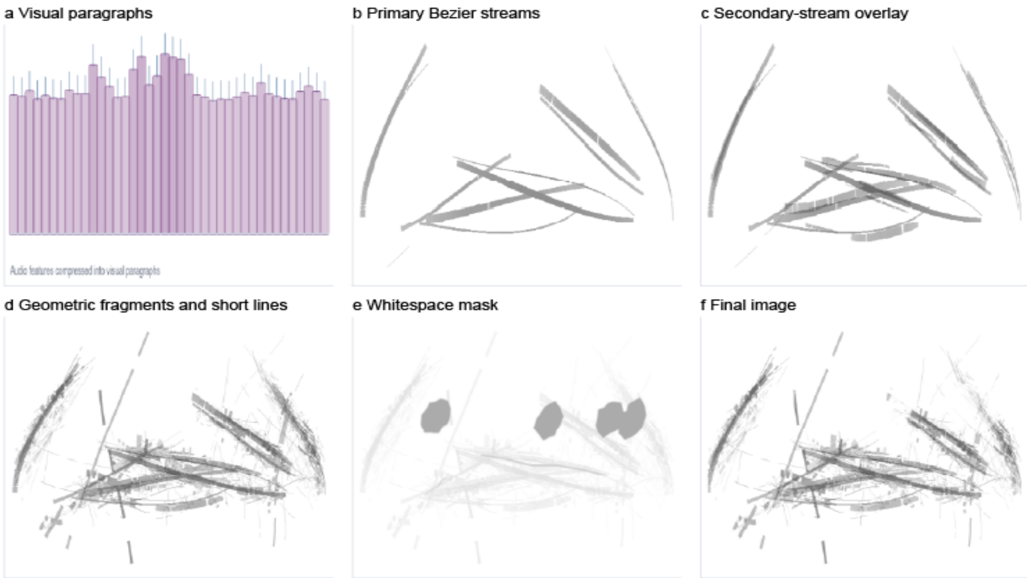

Fig 3. Visual generation mechanism of sound trace flow composition

3.4. Preview and Re-rendering

The system follows a parameter adjustment → low-resolution preview → high-resolution re-rendering workflow. The system first generates a low-resolution preview to provide rapid feedback on composition, density, and local structure. The final image is not obtained by scaling the preview, but is rendered again from the same audio analysis, parameter values, and random seed. This prevents low-resolution sampling artifacts from being carried into the finished work. With a fixed random seed, identical audio and parameter settings reproduce the same overall composition and element distribution, whereas changing the seed generates a new compositional variant while preserving the acoustic constraints and underlying visual grammar.

Sound Trace Flow Composition is therefore a rule-based procedural system in which the visual result is jointly determined by the birdsong input, mapping rules, parameter settings, and random seed. The study aims to explore an interpretable mechanism for generating visual images from audio, with its generation logic understandable to people. The audio-visual correspondences have not yet been independently validated and should therefore be regarded as explicit rules within the artwork rather than universal crossmodal laws.

4. Visual Outcomes

All birdsong samples follow the same feature-extraction and mapping rules, producing images with a consistent visual grammar while preserving differences in duration and spectral structure through changes in flow, density, scale, continuity, and negative space.

Bird appearance, plumage, habitat, and other semantic information are not used. The images are generated solely from the acoustic structure of the recordings.

Controlled randomness creates local variation without altering the core rules. The mapping between the source audio and visual structures remains traceable through the defined rules.

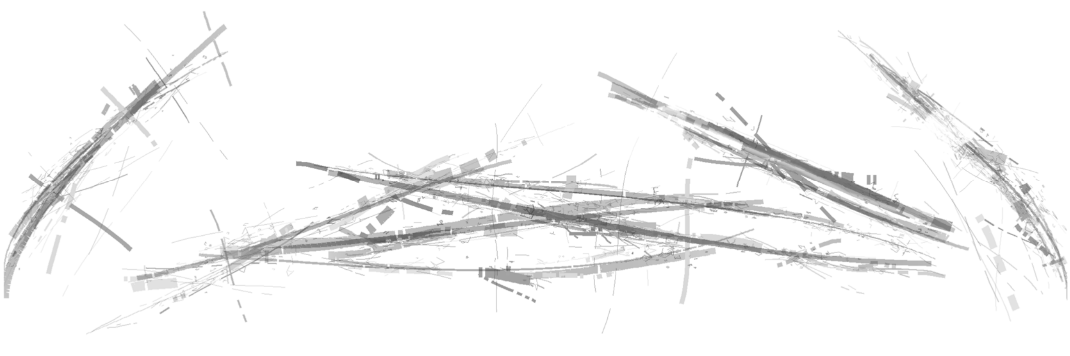

Fig 4. Complete generated long composition

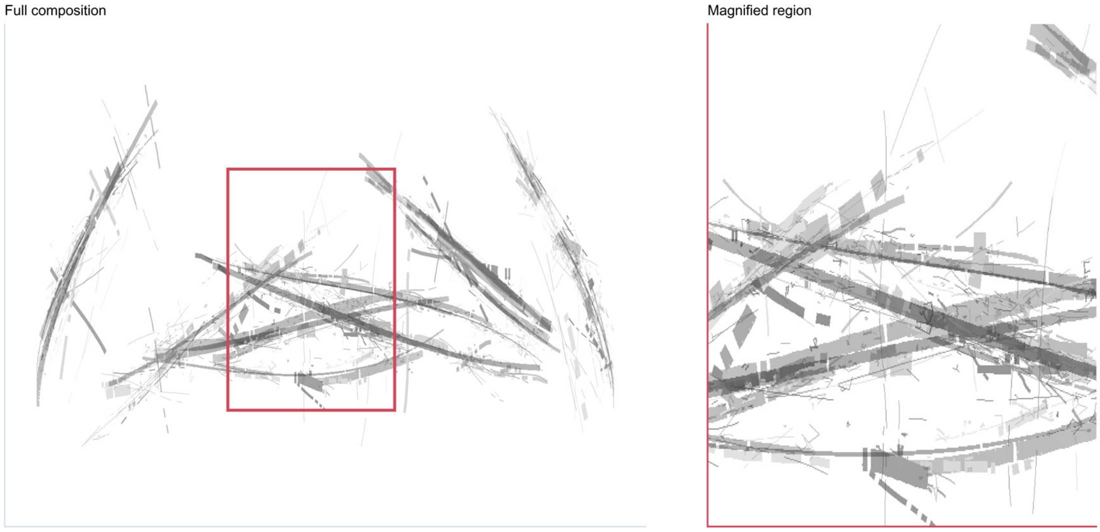

Fig 5. Local magnified detail

5. Conclusion and Discussion

This paper presents Sound Trace Flow Composition, a rule-based method that turns birdsong features into procedural visuals. RMS, spectral centroid, frequency-band energy, and zero-crossing rate control flow lines, fragments, density, continuity, and negative space. A fixed random seed makes the result reproducible.

Different recordings produce different compositions within the same visual grammar. The method stays traceable to the source audio, but it does not claim a universal sound-image correspondence.

The main limits are the lack of perceptual validation, ablation studies, parameter-sensitivity analysis, and viewer evaluation. Future work will test other mappings, examine individual acoustic features, conduct audience research, and extend the method to other natural sounds.

References

[1] Lima, H. B., dos Santos, C. G. R., & Meiguins, B. S. (2021). A survey of music visualization techniques. ACM Computing Surveys, 54(7), 1-29. https://doi.org/10.1145/3461835.

[2] Foote, J. (1999). Visualizing music and audio using self-similarity. Proceedings of the Seventh ACM International Conference on Multimedia, 77-80. https://doi.org/ 10.1145/ 319463. 319472.

[3] Dzwonczyk, L., Cella, C. E., & Ban, D. (2024). Network bending of diffusion models for audio-visual generation. Proceedings of the 27th International Conference on Digital Audio Effects (DAFx24).

[4] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), 293-302. https://doi.org/10.1109/ TSA. 2002. 800560.

[5] Grill, T., & Flexer, A. (2012). Visualization of perceptual qualities in textural sounds. Proceedings of the International Computer Music Conference (ICMC 2012).

[6] Spence, C. (2011). Crossmodal correspondences: A tutorial review. Attention, Perception, & Psychophysics, 73(4), 971-995. https://doi.org/10.3758/s13414-010-0073-7.

[7] Rodrigues, A., Sousa, B., & Cardoso, A. (2022). "Found in translation": An evolutionary framework for auditory-visual relationships. Entropy, 24(12), 1706. https://doi.org/10. 3390/ e24121706.

[8] Shim, J. Y., Kim, J., & Kim, J. K. (2023). Audio-to-visual cross-modal generation of birds. IEEE Access, 11, 27719-27729. https://doi.org/10.1109/ACCESS.2023.3257565.

[9] Kim, S. B., Senocak, A., & Ha, H. (2023). Sound to visual scene generation by audio-to-visual latent alignment. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6430-6440.

[10] Lee, C. C., Lin, W. Y., & Shih, Y. T. (2020). Crossing you in style: Cross-modal style transfer from music to visual arts. Proceedings of the 28th ACM International Conference on Multimedia. https://doi.org/10.1145/3394171.3413624.

[11] Chen, Z., Geng, D., & Owens, A. (2024). Images that sound: Composing images and sounds on a single canvas. Advances in Neural Information Processing Systems 37. https://doi.org/ 10. 52202/079017-2700.

[12] Routray, P. K., Nagpal, M., & Gupta, M. (2025). AI-generated visualizations of musical patterns. ShodhKosh: Journal of Visual and Performing Arts, 6(2s), 168-178. https://doi.org/10. 29121/shodhkosh.v6.i2s.2025.6697.

[13] Patra, B., Jeyanthi, P., & Bhardwaj, R. (2025). Cross-disciplinary art through AI-generated music and visuals. ShodhKosh: Journal of Visual and Performing Arts, 6(4s), 245-254. https://doi.org/10.29121/ shodhkosh. v6.i4s. 2025. 6832.

[14] Liu, Y., Luan, H., & Liu, D. (2026). ChladniSonify: A visual-acoustic mapping method for Chladni patterns in new media art creation [Preprint]. arXiv. https://arxiv.org/ abs/ 2605. 09846.