



Audio Engineering Society

Convention Express Paper 333

Presented at the 158th Convention

2025 May 22-24, Warsaw, Poland

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Exploring Process for Interjoined Procedurally Generated Visual and Audial Content

Grzegorz Samson¹

¹ Feliks Nowowiejski Academy of Music in Bydgoszcz, ul. Słowackiego 7, 85-008, Bydgoszcz, Poland

Correspondence should be addressed to Grzegorz Samson (grzegorzsamson218@hotmail.com)

ABSTRACT

This paper investigates the innovative synthesis of procedurally generated visual and auditory content through the use of Artificial Intelligence (AI) Tools, specifically focusing on Generative Pre-Trained Transformer (GPT) networks. This research explores the process of procedurally generating an audiovisual representations of semantic context by generating images, artificially providing motion and generating corresponding multilayered sound. The process enables the generation of stopped-motion audiovisual representations of concepts.

This approach not only highlights the capacity for Generative AI to produce cohesive and semantically rich audiovisual media but also delves into the interconnections between visual art, music, sonification, and computational creativity. By examining the synergy between generated imagery and corresponding soundscapes, this research paper aims to uncover new insights into the aesthetic and technical implications of the use of AI in art.

This research embodies a direct application of AI technology across multiple disciplines creating intermodal media. Research findings propose a novel framework for understanding and advancing the use of AI in the creative processes, suggesting potential pathways for future interdisciplinary research and artistic expression. Through this work, this study contributes to the broader discourse on the role of AI in enhancing creative practices, offering perspectives on how various modes of semantic representation can be interleaved using state-of-the-art technology.

1 Introduction

The rapid advancement of contemporary transformer-based neural networks continuously introduces new opportunities and research directions for both academics and wide group of web users [1], [2]. Increasingly sophisticated forms of generated content — spanning textual, visual, and auditory domains — are becoming prevalent within the digital space. Internet creators have begun utilizing transformer networks for generating diverse types of media, including visual media [3], synthesized speech, musical media, trying to integrate these modalities into an audiovisual experiences [4].

In this evolving landscape, the role of the user is shifting toward that of an intermediary — someone who mediates, curates, and integrates various types of generated media [5]. As the range of AI-generated tools and content expands, a crucial question emerges: how can procedural generation process be developed to produce meaningfully composed audiovisual works? Such an approach which could systematically unify visual and auditory outputs through automated yet semantically interconnected generative methods, encouraging further exploration into the interplay between computational creativity,

multimodal representation, and human-guided mediation.

2 Existing Solutions

2.1 Audio Generation

The field of audio generation has seen rapid advancements fuelled by machine learning technologies [6], particularly in speech synthesis [7], [8], [9]. Another significant area of progress is the synthesis of musical content [10], [11], [12]. Moreover, a third prominent category comprises models designed for the general generation of audio content, such as sound effects [12], [13], [14], [15], [16]. Open-source implementations of such models have been introduced by industry leaders including Meta [17]. One of the prominent providers of such consumer-oriented audio generation technologies is Eleven Labs [18], which offers sophisticated tools widely used across creative and commercial contexts.

2.2 Image Generation

Visual content remains one of the most prominent outputs within contemporary generative artificial intelligence (GenAI). Early explorations, exemplified by the experimental work of Google's DeepDream laboratory [19], laid foundational groundwork for subsequent advancements [20]. This trajectory accelerated significantly with the introduction of influential publicly accessible tools, such as Stable Diffusion [21], [22], and artistically acclaimed iterations of Midjourney models [23], [24] and OpenAI Dall-E [25], [26].

2.3 Video Generation

A significant milestone in the evolution of generative technologies has been the emergence of models capable of synthesizing coherent sequences of related image frames, thereby enabling the creation of dynamic video content [27], [28]. Prominent examples include platforms such as OpenAI's Sora [29] and Runway's Gen-2 [30].

Despite these advancements, video content produced by generative AI varies in quality and remains predominantly limited to short-form outputs. This is largely due to the high computational costs associated with generating longer sequences, as well as the technical challenges that models face in maintaining semantic coherence and visual consistency over extended durations [31].

2.4 Video with Audio Integration

Artists explore the possibilities of intertwining different modalities [32], with generative models increasingly capable of producing multimodal narratives from simple textual prompts [33]. These models can generate visual sequences accompanied by voice-over narration or soundtracks.

Recent developments indicate a shift toward audio generation frameworks which use a video as input. For instance, Eleven Labs has developed a prototype system implemented in Node.js, which leverages AI-based image recognition to analyze selected frames (4 frames for every second) within a video [34]. Based on this analysis, the system generates corresponding audio proposals using sound effect (SFX) generation models. Despite this advancement, the audio output remains limited in scope: typically consisting of a single-layered audio track. This constraint restricts the potential for nuanced auditory design, particularly in terms of spatiality and layering.

2.5 Procedural Audiovisual Generation

The solutions discussed so far enable the creation of audiovisual works that are finite in duration and often accompanied by a limited, static audio layer. This research paper aims to explore a different process: one in which a single source input is transformed into both a visual content, taking the form of a looped, dynamic video, referred to as “image-in-motion”, and a generative audial layer. These two media streams would coexist side by side, remain interdependent, and maintain a semantic relationship, yet without being constrained by a fixed duration.

3 Procedural Pipeline

3.1 Strategies

In developing procedural methods for generating semantically linked visual and auditory media, various strategies can be employed depending on the intended level of interconnection between the two modalities.

In a basic approach, a text input is used to guide a transformer-based system that generates both visual and audio layers. The initial text input can be improved through additional processing to make the results more coherent and relevant. The following sections describe specific strategies used to enhance and coordinate the generation of both visual and audio elements.

3.1.1 Visual Modality

- **Prompt Derivation:** A dedicated visual prompt is generated from the textual source input, optimized for image generation models.
- **Video Synthesis from Reference Imagery:** A static reference image is created based on the prompt, which is then animated to produce a short video clip.

3.1.2 Auditory Modality

- **Prompt Derivation:** A separate text prompt is crafted for the audio layer, derived from the same source input, tailored to sound generation models.
- **Layered Sound Design:** The audio description is decomposed into multiple sound elements or layers, enabling the generation of distinct audio components that together form a richer and more complex soundscape.
- **Parametric Playback:** The multi-layered soundscape is designed to be played in a randomized and parameterized fashion, introducing variation and dynamism during playback.

3.1.3 Semantic Alignment Across Modalities

- **Joint Prompting:** A single text prompt is created to guide both the visual and auditory models simultaneously, ensuring initial semantic consistency between the two outputs.
- **Sequential Prompting:** A visual prompt is generated first, followed by an auditory prompt derived from the visual output, or vice versa, using image, or sound recognition.
- **Correlated Prompting:** Two separate, but semantically related, text prompts are generated — one for the visual model and one for the auditory model. These prompts are derived from the same source input but are tailored to suit the specific needs of each modality, maintaining thematic and coherence across both outputs.

3.1.4 Output Variability

The strategies outlined above offer a structured approach to aligning visual and auditory elements within a procedural generation system. However, it is important to recognize the inherent unpredictability of transformer-based models. Each generation can vary in terms of semantic coherence, with the quality

and consistency of the output heavily dependent on the nature and clarity of the input prompts. Maintaining semantic alignment between the generated audio and visual components is essential to ensure that both modalities contribute meaningfully to a unified, coherent compound media.

4 Practical Implementation

4.1 Technologies Used

To construct a procedural pipeline capable of generating semantically linked audio and visual media, publicly available models running on local user machines can be used. However, generating high-quality outputs — especially videos — locally requires substantial computational resources, often limiting accessibility. As an alternative, cloud-based commercial services provide hosted generative solutions that offer superior performance and output quality.

For content processing within transformer-based models, commercially available solutions accessible via consumer markets and robust developer-oriented APIs were employed. Specifically, the prototype utilized OpenAI's *GPT-4o-mini* for textual processing, and *DALL-E 3* for visual generation, *Runway Gen-2 (Gen4_turbo)* for video synthesis, and Eleven Labs *text to sound effects* model for audio generation.

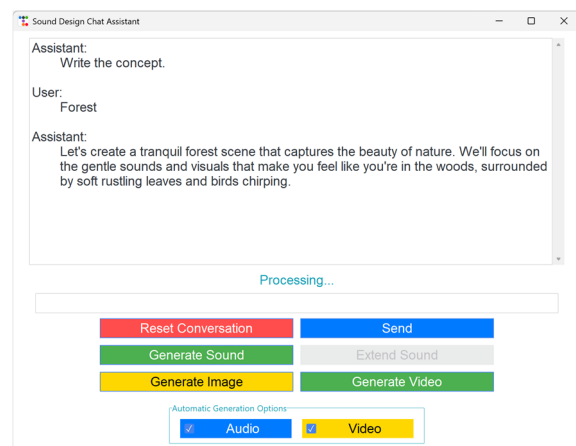


Figure 1. Python App GUI. (source: own research)

4.2 GUI Application

A prototype Python application was developed to demonstrate the procedural pipeline for generating semantically interconnected audio and visual layers. The application communicates with a custom

granular synthesizer, implemented as a VST plugin, which runs independently within the Reaper digital audio workstation (DAW).

To facilitate both prototyping and comprehensive automated testing, a functional graphical user interface (GUI) was developed, allowing interactive experimentation through user input (as shown in Fig. 1). The system architecture employs a virtual assistant functioning as a middleware. This assistant interprets user messages, converting them into structured instructions for generating the associated audio and visual layers.

Source application code is accessible at Open Science Framework [35].

4.3 Application Logic

The application operates based on the following assumptions:

- User-provided source text serves as the initial input.
- The source text is processed into separate prompts for audio and visual generation.
- These prompts are sent to the respective transformer-based models for audio and visual content generation.
- The resulting visual output is displayed, while the audio output is played back through the granular synthesizer.

The assistant module is controlled by a complex instruction prompt defined in the “instructions.yaml” file. This instruction prompt specifies both the content generation methods and formatting guidelines (JSON format is used for model responses). This structured communication enables controlled model responses while preserving content variability.

The assistant generates the following outputs:

- Conversational message
- Visual text prompt
- Sound text prompt
- System action (for automatic generation)

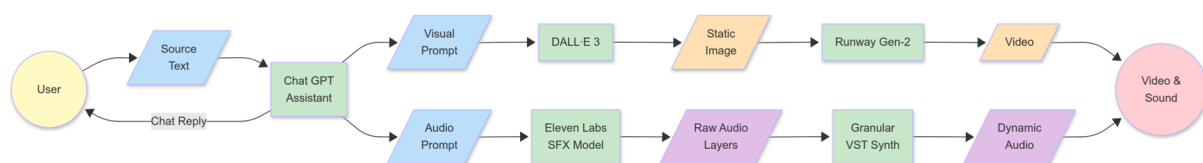


Figure 2. Procedural Audiovisual Content Generation Pipeline. (source: own research)

Each prompt is tailored to the specific transformer model involved.

Within the video generation process:

- An image is generated based on the text prompt.
- The image is saved locally.
- The image is resized and uploaded to the IMGUR server to generate an embeddable link.
- Both the image link and the original text prompt are submitted to the video generation model.

The overall process is illustrated in Fig. 2. All generated media are archived in a separate directory, with dedicated tags and filenames derived from the assistant's response JSON, enabling efficient organization, identification, and retrieval.

4.4 Generative Sound Projection

To achieve the project's requirement for dynamic audio playback that maintains both continuity and variability, a custom granular synthesizer was developed as a VST plugin (as shown in Fig. 3). This granular synthesizer continuously generates audio content by randomly selecting and playing short audio grains from a predefined collection of sound files. It was specifically designed to temporally extend relatively short audio samples (typically 10–20 seconds long) generated by Eleven Labs.

The synthesizer, inspired by Anthony Brigante's open-source project [36], was adapted into a VST plugin and enhanced to support multiple audio files. It also features a dynamic “watch folder” function, which automatically loads new audio files added to the monitored directory. This integration ensures that user-generated prompts dynamically trigger corresponding audiovisual outputs within the procedural pipeline. Source code for the VST plugin is accessible at Open Science Framework [37].

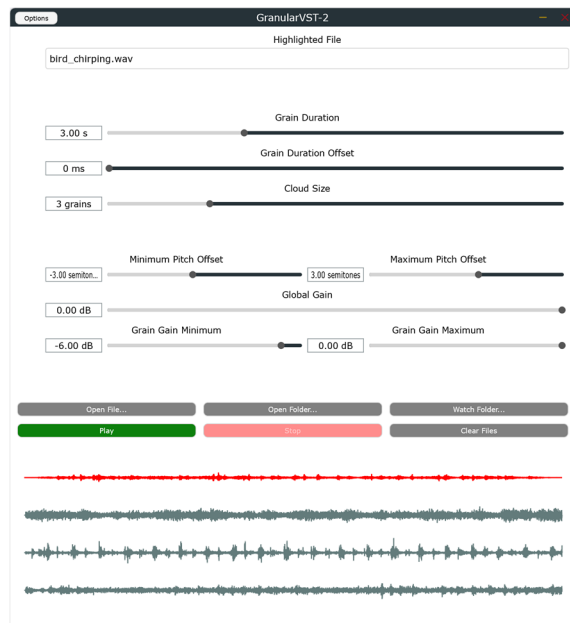


Figure 3. Custom Granular Synthesizer VST GUI. (source: own research)

5 Results

Preliminary empirical testing has revealed potential trends, opportunities, and challenges. The prototype application successfully represents both realistic and diverse abstract concepts, demonstrating coherent audiovisual media.

5.1 Highlights

The use of an assistant capable of generating semantically linked text prompts for both visualization and sonification enabled a consistent interpretive vision based on the given text input. For example, in response to emotion-based prompts such as *anger*, *sorrow*, or *anticipation*, the assistant generated visual and auditory representations that captured the essence of each emotion. For *anger*, the system produced an image of a storm, accompanied by sounds such as “distant thunder,” “howling wind,” “scraping stones,” and “crack of lightning”.

5.2 Challenges

The proposed solution also encounters several challenges. One critical limitation lies in the image-to-video transformation process. The model often struggles to animate complex scenes with multiple objects accurately. For instance, attempts to generate animated sequences involving penguins or dragons resulted in a simplistic “zoom-in” effect rather than realistic motion. In the case of a storm, animation was

limited to lightning flashes, lacking dynamic elements such as wind gusts.

The audio layer, while functional, is not always compelling from an auditory perspective—particularly when generating ambient or noise-based soundscapes, which tend to lack richness and variation.

It is important to emphasize that the tested prototype allows for decoupling of the individual generation stages, thereby creating opportunities for user intervention. Human verification and refinement can undoubtedly enhance the overall quality and semantic coherence of the generated media.

6 Conclusions

The proposed solution presents promising opportunities, serving as a practical demonstration of the procedural generation of interjoined audiovisual content. At its current stage, this solution remains a prototype undergoing initial testing and validation. The underlying technology should be the subject of extensive exploration and experimentation within both academic research and creative practice.

6.1 Future Research

Future research directions could include optimizing the coherence and dynamism of generative outputs, refining the integration mechanisms between visual and auditory modalities, and improving control mechanisms for human interaction within the procedural workflow. Additionally, investigating user perception and reception of procedurally generated media would provide valuable insights into its potential applications across various fields, including entertainment, immersive environments, interactive installations, and therapeutic contexts.

7 Acknowledgements

This work was carried out within the supportive environment of the Feliks Nowowiejski Academy of Music in Bydgoszcz, an institution dedicated to fostering innovative and exploratory research. Special thanks are extended to Rector Prof. Elżbieta Wtorkowska and Dean Prof. Aleksandra Kłaput-Wiśniewska, and Dr Katarzyna Figat. Special appreciation goes to Prof. Barbara Okoń-Makowska and Dr Krzysztof Cybulski for their encouragement to explore broader perspectives on generative media.

References

- [1] A. Vaswani *et al.*, “Attention Is All You Need,” Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
- [2] S. Bubeck *et al.*, “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” *arXiv.org*, 2023, doi: 10.48550/arXiv.2303.12712.
- [3] H. Matatov, M. A. L. Qu’er’e, O. Amir, and M. Naaman, “Examining the Prevalence and Dynamics of AI-Generated Media in Art Subreddits,” *arXiv.org*, 2024, doi: 10.48550/arxiv.2410.07302.
- [4] Y. Lyu, H. Zhang, S. Niu, and J. Cai, “A Preliminary Exploration of YouTubers’ Use of Generative-AI in Content Creation,” in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, in CHI EA ’24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–7. doi: 10.1145/3613905.3651057.
- [5] J. Edwards, A. Perrone, and P. R. Doyle, “Transparency in Language Generation: Levels of Automation,” *CIU*, Jun. 2020, doi: 10.1145/3405755.3406136.
- [6] D. Ș. Rucăreanu and L. Lăzărescu-Thois, “The Impact of AI in the Field of Sound for Picture. A Historical, Practical, and Ethical Consideration,” *CONCEPT*, vol. 29, no. 2, Art. no. 2, Dec. 2024, doi: 10.37130/bf780v59.
- [7] C. Zhang, S. Zheng, M. Zhang, M. Qamar, S. Bae, and I. S. Kweon, “A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI,” *arXiv.org*, Mar. 2023, doi: 10.48550/arxiv.2303.13336.
- [8] P. K. Rubenstein *et al.*, “AudioPaLM: A Large Language Model That Can Speak and Listen,” *arXiv.org*, Jun. 2023, doi: 10.48550/arxiv.2306.12925.
- [9] M. Le *et al.*, “Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale,” *Neural Inf. Process. Syst.*, Jun. 2023, doi: 10.48550/arxiv.2306.15687.
- [10] A. Agostinelli *et al.*, “MusicLM: Generating Music From Text,” *arXiv.org*, Jan. 2023, doi: 10.48550/arxiv.2301.11325.
- [11] Q. Huang *et al.*, “Noise2Music: Text-conditioned Music Generation with Diffusion Models,” Mar. 06, 2023, *arXiv*: arXiv:2302.03917. doi: 10.48550/arXiv.2302.03917.
- [12] A. Vyas *et al.*, “Audiobox: Unified Audio Generation with Natural Language Prompts,” Dec. 25, 2023, *arXiv*: arXiv:2312.15821. doi: 10.48550/arXiv.2312.15821.
- [13] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” *Speech Synth. Workshop*, p. 125, Sep. 2016, doi: 10.48550/arXiv.1609.03499.
- [14] H. Liu *et al.*, “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models,” *Int. Conf. Mach. Learn.*, Jan. 2023, doi: 10.48550/arxiv.2301.12503.
- [15] Z. Evans, J. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, “Stable Audio Open,” *arXiv.org*, 2024, doi: 10.48550/arxiv.2407.14358.
- [16] J. Liang *et al.*, “WavCraft: Audio Editing and Generation with Large Language Models,” May 10, 2024, *arXiv*: arXiv:2403.09527. doi: 10.48550/arXiv.2403.09527.
- [17] F. Kreuk *et al.*, “AudioGen: Textually Guided Audio Generation,” Mar. 05, 2023, *arXiv*: arXiv:2209.15352. doi: 10.48550/arXiv.2209.15352.
- [18] “Sound effects | ElevenLabs Documentation.” Accessed: Apr. 27, 2025. [Online]. Available: <https://elevenlabs.io/docs/capabilities/sound-effects>
- [19] A. Mordvintsev, C. Olah, and M. Tyka, “Inceptionism: Going Deeper into Neural Networks.” 2015. [Online]. Available: <https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>
- [20] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” Dec. 16, 2020, *arXiv*: arXiv:2006.11239. doi: 10.48550

- /arXiv.2006.11239.
- [21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” *Comput. Vis. Pattern Recognit.*, Jun. 2022, doi: 10.1109/cvpr52688.2022.01042.
- [22] P. Esser *et al.*, “Scaling Rectified Flow Transformers for High-Resolution Image Synthesis,” Mar. 05, 2024, *arXiv*: arXiv:2403.03206. Accessed: Nov. 12, 2024. [Online]. Available: <http://arxiv.org/abs/2403.03206>
- [23] “Midjourney.” Accessed: Apr. 27, 2025. [Online]. Available: <https://docs.midjourney.com>
- [24] D. Hanna, “The Use of Artificial Intelligence Art Generator ‘Midjourney’ in Artistic and Advertising Creativity,” *J. Des. Sci. Appl. Arts*, vol. 4, no. 2, pp. 42–58, Jun. 2023, doi: 10.21608/jdsaa.2023.169144.1231.
- [25] A. Ramesh *et al.*, “Zero-Shot Text-to-Image Generation,” *Int. Conf. Mach. Learn.*, pp. 8821–8831, Jul. 2021, doi: 10.48550/arXiv.2102.12092.
- [26] A. Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, “Hierarchical Text-Conditional Image Generation with CLIP Latents,” *arXiv.org*, 2022, doi: 10.48550/arxiv.2204.06125.
- [27] “Meta Movie Gen.” Accessed: Nov. 12, 2024. [Online]. Available: <https://ai.meta.com/research/movie-gen/>
- [28] D. Kondratyuk *et al.*, “VideoPoet: A Large Language Model for Zero-Shot Video Generation” doi: 10.48550/arXiv.2312.14125.
- [29] Y. Liu *et al.*, “Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models,” Apr. 17, 2024, *arXiv*: arXiv:2402.17177. doi: 10.48550/arXiv.2402.17177.
- [30] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, “Structure and Content-Guided Video Synthesis with Diffusion Models,” Feb. 06, 2023, *arXiv*: arXiv:2302.03011. doi: 10.48550/arXiv.2302.03011.
- [31] C. Li, D. Huang, Z. Lu, Y. Xiao, Q. Pei, and L. Bai, “A Survey on Long Video Generation: Challenges, Methods, and Prospects,” Mar. 25, 2024, *arXiv*: arXiv:2403.16407. doi: 10.48550/arXiv.2403.16407.
- [32] A. Ploin, R. Eynon, I. Hjorth, and M. A. Osborne, “AI and the Arts: How Machine Learning is Changing Artistic Work,” Oxford Internet Institute, University of Oxford, UK, Report, 2022.
- [33] L. G. Foo, H. Rahmani, and J. Liu, “AI-Generated Content (AIGC) for Various Data Modalities: A Survey,” Jan. 19, 2025, *arXiv*: arXiv:2308.14177. doi: 10.48550/arXiv.2308.14177.
- [34] “elevenlabs-examples/examples/sound-effects/video-to-sfx at main · elevenlabs/elevenlabs-examples.” Accessed: Apr. 24, 2025. [Online]. Available: <https://github.com/elevenlabs/elevenlabs-examples/tree/main/examples/sound-effects/video-to-sfx>
- [35] G. Samson, “Procedurally Generated Visual and Audial Content App,” Apr. 2025, Accessed: Apr. 25, 2025. [Online]. Available: <https://osf.io/e65n8/>
- [36] A. Brigante, *GranularSynth*. (Dec. 19, 2024). C++. Accessed: Apr. 27, 2025. [Online]. Available: <https://github.com/abrigante1/GranularSynth>
- [37] G. Samson, “Procedurally Generated Visual and Audial Content Granular VST,” Apr. 2025, Accessed: Apr. 27, 2025. [Online]. Available: <https://osf.io/9m5dh/>